Technology Service Corporation

2811 WILSHIRE BOULEVARD • SANTA MONICA, CALIFORNIA 90403 • PH. (213) 829-7411

DEVELOPMENT OF IMPROVED METHODS FOR PREDICTING AIR QUALITY LEVELS IN THE SOUTH COAST AIR BASIN

Final Report

March 1979

by

Melvin D. Zeldin
Joseph C. Cassmassi

TSC-PD-B572-10

Submitted to: California Air Resources Board
Sacramento, California 95814

Mr. Charles Bennett, Project Officer
Contract Number A6-192-30

DISCLAIMER

The statements and conclusions in this report are those of the Contractor and not necessarily those of the California Air Resources Board. The mention of commercial products, their source or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products.

This report is submitted in fulfillment of Contract Number A6-192-30 by Technology Service Corporation, Santa Monica, California, under the sponsorship of the California Air Resources Board. Work was completed in November, 1978.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

Chapter Page

TABLE OF CONTENTS (cont'd)

LIST OF FIGURES

LIST OF FIGURES (cont'd)

LIST OF FIGURES (cont'd)

LIST OF TABLES

LIST OF TABLES (cont'd)

LIST OF TABLES (cont'd)

ACKNOWLEDGMENTS

The authors wish to thank the many people who helped to make this report possible. The comments and suggestions by Charles Bennett, our ARB project officer, and John Kinney, ARB senior meteorologist, were particularly helpful. The assistance in time series analysis was provided by Dr. Kent Wilson of the University of California, San Diego.

At Technology Service Corporation, technical guidance by Drs. John Trijonis, Leo Breiman, and Larry Lai, and computer support by Dennis Bicker, George Berglund, Louis Gurtowski, and Frank Nakatsuma are greatly appreciated. Finally, special thanks are due to Pat Robertson and Nina Stivers for their assistance in the preparation of this report.

CHAPTER 1
INTRODUCTION

The State of California's Air Pollution Emergency Plan outlines actions to be taken when air pollution levels reach or are expected to reach specified episode levels. The emergency actions include protective measures (such as health warnings) as well as preventive measures (such as industrial emission abatement programs).

Since the episode concentration levels (see Table 1.1) imply the probable existence of health-threatening conditions, it is highly desirable to predict these episodes in advance, so that protective and/or preventive action can be taken with sufficient lead time. The expense, disruption, and credibility of these protective/preventive actions requires a minimization of false-alarm rate. Thus, there is a need for accurate air quality forecasting in support of the Air Pollution Emergency Plan.

Under contract to the California Air Resources Board (ARB), Technology Service Corporation has developed improved prediction techniques in the South Coast Air Basin (SCAB) for three specific pollutants: oxidants, sulfates, and sulfur dioxide.

1.1 OBJECTIVES OF THE PROJECT

While the primary objective of this research is to develop improved prediction algorithms for specified sites in the SCAB, other preliminary requirements were necessary. These are summarized below:

(1) Review the Existing State-of-the-Art

An extensive literature review of prediction methods was undertaken, including both meteorological and pollutant relationships. Further, existing methods for predicting pollutant levels in the SCAB were reviewed, including objective and subjective methods by both the ARB and the South Coast Air Quality Management District (AQMD), and stagnation advisory criteria used by the National Weather Service Forecast Office in Los Angeles (WSFO).

(2) Development of Verification Methods

In order to establish a baseline accuracy of existing predic-

Table 1.1 Appropriate ARB Episode Criteria

Air Contaminant	Averaging Time	Episode Criteria		
		Stage 1	Stage 2	Stage 3
Photochemical Oxidant (including Ozone)	1 Hour	0.20 ppm	0.35 ppm	0.50 ppm
Sulfur Dioxide	1 Hour	0.50 ppm	1.00 ppm	2.00 ppm
	24 Hours	0.20 ppm	0.70 ppm	0.90 ppm
Oxidant, in Combination with Sulfur Dioxide*	1 Hour	0.20 ppm	0.35 ppm	0.50 ppm
Sulfate, in Combination with Oxidant	24 Hours (Sulfate)	25 $\mu g/m^3$		
	1 Hour (Oxidant)	0.20 ppm		

* Both oxidant and sulfur dioxide must be greater than 0.10 ppm.

ABSTRACT

This is the Final Report of a project to develop improved prediction algorithms for oxidants, sulfates, and sulfur dioxide in the South Coast Air Basin (SCAB). The main objectives of the study included: (1) a comprehensive survey of existing prediction methods, (2) a review of existing pollution prediction methods used by the Air Resources Board (ARB), the South Coast Air Quality Management District (AQMD), and the National Weather Service Forecast Office in Los Angeles (WSFO), (3) the development of a method to evaluate and verify prediction algorithms, (4) the compilation of 1974-1977 aerometric data (with 1974-1976 to be used as a dependant data set, and 1977 to be used as an independant test set), and (5) the development of new prediction algorithms for same-day and day-in-advance application.

A Phase I report, completed in October 1977, described the existing state-of-the art and presented a method for evaluating prediction techniques. The current report summarizes the earlier effort and presents the results and verification of newly developed algorithms. For each of the three specified pollutants, key prediction sites were determined. Algorithms were developed separately for each of these sites and then related to the remaining SCAB sites by means of linear regression. Thus it is possible to predict the level of a specific pollutant at each SCAB site currently measuring it.

The major findings of the study are as follows: (1) the new algorithms have substantially improved the capability for same-day predictions, and to a lesser degree for day-in-advance prediction; (2) the statistical relationships between meteorology and pollutant concentrations degrade rapidly as the lead-time increases, such that historical data are ineffective for 30-hour predictions; (3) the use of numerical prognostic charts (LFM progs) have improved day-in-advance prediction methods; and (4) for SO_2, the inclusion of emission factors does not appreciably improve prediction accuracies over those obtained from meteorological data alone.

tion methods, an evaluation technique was developed to incorporate four major aspects of pollution prediction:

(a) episode category accuracy
(b) quantitative accuracy
(c) significant change accuracy
(d) reduction of episode false alarms while increasing episode probability of detection

Baseline accuracies were determined for Upland and Downtown Los Angeles, from which future algorithms could be compared.

(3) Compilation of Data Bases

An extensive data base comprising surface and upper air meteorological variables over a four-state region, site specific pollutant concentrations, and power plant sulfur dioxide emissions were compiled for the 1974 through 1976 period. In all, over 350 potential predictor variables were assembled. For independent data testing, selected subsets of these variables for the 1977 year were also collected.

(4) Development of Improved Prediction Algorithms

Specifically, algorithms were to be developed for key sites in the SCAB and for definitive prediction lead times. For oxidant and sulfur dioxide predictions, lead times of 4-10 hours (same-day) and 24-30 hours (day-in-advance) were required. For sulfates, a lead time of 24-hours was necessary. Key sites selected included: (1) oxidant: Upland, Downtown Los Angeles, La Habra, Riverside, and Newhall; (2) sulfates: Anaheim, Azusa, Riverside, Reseda, Temple City, and Upland; and (3) sulfur dioxide: Lennox and Fontana. In addition, prediction equations were to be developed relating the oxidant and sulfur dioxide key sites to the remaining SCAB locations currently monitored on the ARB telemetry system (see Figure 1.1).

1.2 SUMMARY OF RESULTS

Based on the research effort conducted in this project, the following is a summary of the principal results:

- Same-Day Prediction

Significant improvement was made over existing best-available

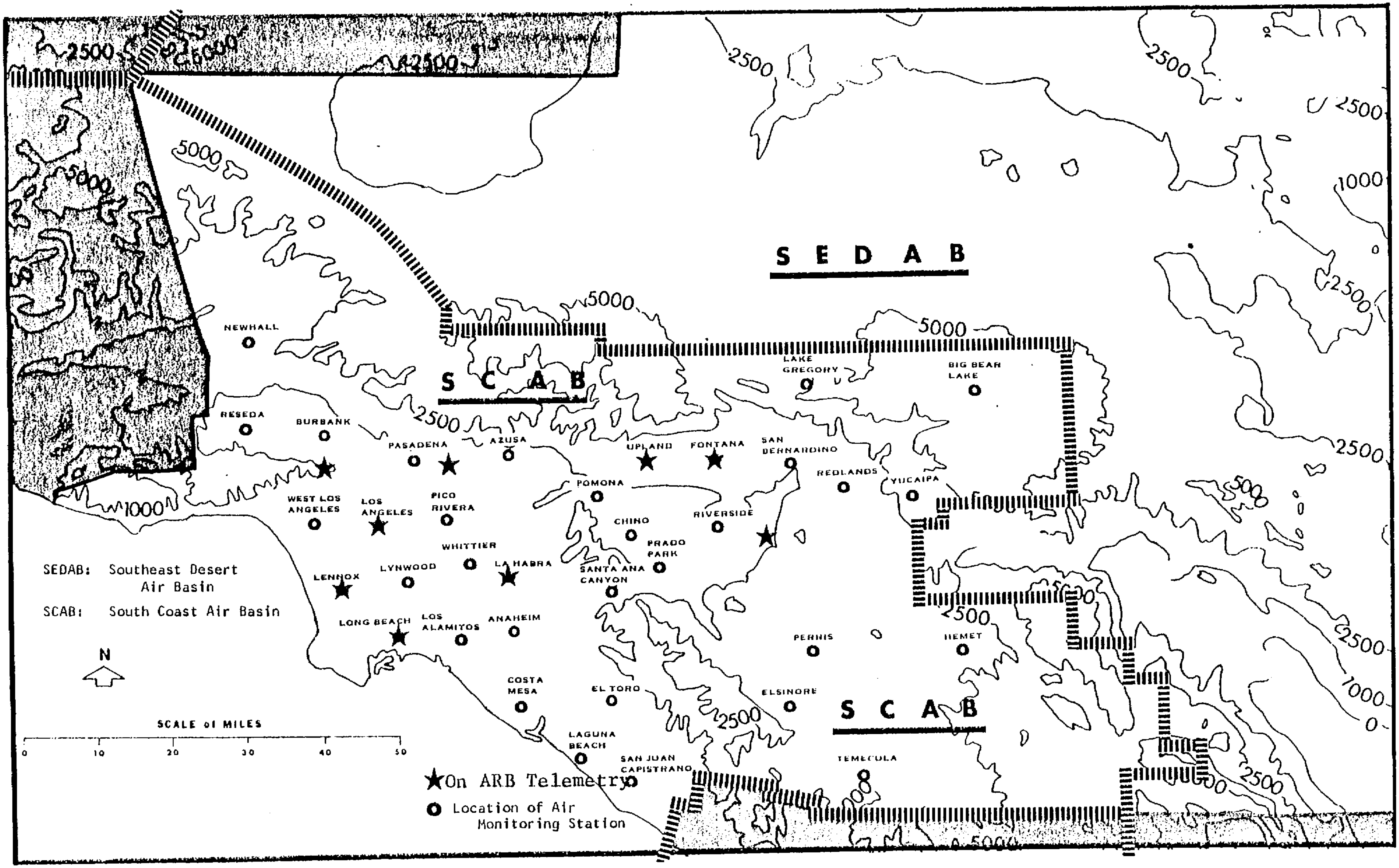


Figure 1.1 Location of Air Monitoring Sites in the SCAB

prediction algorithms for oxidant. For each of the five key sites, improvement occurred on both the dependent and independent data sets. Interestingly, the most successful algorithms for Upland, Riverside, Newhall, and La Habra were developed interactively (i.e., human expertise modifying computerized statistical analyses). These methods provided the greatest resolution in predicting both high and low pollution values. The following paragraphs list some of the statistical techniques applied and the probable reason for not performing as well as the interactive methods.

(1) Linear regression (stepwise): Produced a "best fit" of the data, but underpredicted the high concentrations, thus reducing the potential for accurately predicting the most severe situations: Applications of weighted regression (placing emphasis on the high end) produced a slightly better fit at the high end, but over-predicted moderate and low values, thereby increasing the standard error over a non-weighted regression.

(2) Pattern recognition - (AID program): The major difficulty was that only discrete values can be predicted, as opposed to a "continuous" prediction function. The other problem is that the scatter in each prediction terminal node was sufficiently large to significantly impair the verification. Additionally, test sets applied to the AID trees showed that only a small percentage of the variance (25 to 30%) could actually be explained through such techniques.

(3) AID + Regression: Provided little additional reduction of variance. We applied these two ways: (1) to form an AID tree and apply linear regression to the data in the most significant terminal nodes, and (2) to perform regression and then apply AID to the residuals. In both situations, the combined effects increased the explained variance approximately 2-5% over the best individual method. Also, the added complexity as a usable product did not seem appropriate for the very small improvement obtained.

(4) Nearest Neighbor: Did not provide the resolution necessary to be a meaningful prediction method. The possible combinations of meteorological conditions producing similar oxidant levels is quite large. Thus, there was a considerable spread of observed values under matching

meteorological conditions. Results indicated that this method yielded prediction values which approximated climatological values.

(5) Time Series: Provided reduction over persistence of the mean absolute error, but did not successfully predict high concentrations or significant change conditions. One might categorize the Box-Jenkins approach as "enhanced persistence," in that overall results are better than persistence, but it suffers from the same time-lag problems as persistence. Applications to both pollutant values and key meteorological variables failed to yield substantive improvement over other methods.

For SO_2 prediction, persistence is a good technique to use. With rather complicated combinations of meteorological parameters, new algorithms were developed which substantially improved over persistence, thereby providing the ARB with SO_2 predictive capability not previously available. Further a case-study involving daily SO_2 emissions from major power plant point sources near Long Beach, indicated that the addition of such data did not appreciably improve predictive capabilities over those obtained from meteorological data alone.

- Day-in-Advance Prediction

For all three pollutants, prediction algorithms improved over existing methods, but to a lesser degree than for same-day methods. For sulfate prediction, original ARB equations verified quite well. Thus, rather than initiating new algorithms, modifications were made to the existing equations to account for systematic errors. For three critical sites (Upland, Riverside and Reseda), modifications significantly improved predictive capabilities.

Developing algorithms for oxidant prediction was more complex. In general, the use of statistical techniques to relate meteorological conditions to oxidant concentrations was effective within a short time period. However, as the desired lead time increased, the statistical relationships became less pronounced, such that the use of historical data to predict oxidant levels 30 hours in advance was not effective. The principal reason, obviously, can be attributed to a variety of dynamic changes taking place in the atmosphere from one day to the next. In essence, given some initial set of conditions, the end product of oxidant levels on the following day can be quite varied. It is our opinion that 24-hour prediction is the maximum lead time in which historical data can be utilized successfully. Beyond that time period, the degradation in the statistical relationships reduces the end product to approximately persistence and climatology.

Predictive data available from numerical forecasts (progs) issued daily by the National Meteorological Center (NMC) were used for 30-hour day-in-advance pollution prediction. From the Limited Fine Mesh (LFM) prog package, regression equations relating oxidant to predicted 500 mb height values were developed. These prediction equations ("perfect prog") provided better results, in some instances, than the best 24-hour statistical algorithms using historical data. For the first time, therefore, NMC output can be used in objective methods to predict site specific pollution levels.

1.3 ORGANIZATION OF THE REPORT

This report is organized into 8 Chapters. This Chapter provides an introduction and background. Chapter 2 highlights the results of the Phase I report, such that the final report can be used as a stand-alone document. Chapters 3, 4, and 5 describe the algorithms developmental procedures, and verification scores for oxidant, sulfate, and sulfur dioxide, respectively. Methods for estimating missing data are contained in Chapter 6, while Chapter 7 describes a procedure for correcting algorithm output for long-term pollution trend changes. The references are listed in Chapter 8.

CHAPTER 2
REVIEW OF PHASE I

The Phase I report, completed in October 1977, included a comprehensive review of previous and existing predictive capabilities, an evaluation of those methods, and a description of the available data base. This chapter is intended to briefly summarize the highlights of that report in order to provide a sufficient background leading to the results of this research effort. If more detailed background information is desired, the reader is referred to that report (Zeldin and Cassmassi, 1977). Also, parameter abbreviations, not explained in the text, can be found in Appendix A.

2.1 HISTORICAL BACKGROUND

Much of the work to date on oxidant forecasting has been most relevant to the shorter forecasts. Three basic types of approaches have been attempted for making predictions on the short-term scale: time series, multiple regression, and pattern recognition. For an introduction to the field, the reader is encouraged to refer to the printed transcript of the Conference on Forecasting Air Pollution held in Berkeley in 1974 (D. R. Brillinger and E. L. Scott, 1975).

Time series predictions have been studied by McCollister and Wilson (1975), Box and Tiao (1975), and Chock, Levitt, and Terrell (1975). Using univariate time series, the objective is to predict future values based on just the previous values of the time series.

Some results of the work by McCollister and Wilson indicate that time series prediction does a bit better than persistence (that the future mimics the past exactly) and even a bit better than predictions made by trained meteorologists using both meteorological and pollutant data. All three of the methods, however, have fairly large average errors--in the 35-50% range--perhaps too large for actual health warning or short-term emissions control usage.

Chock, Levitt and Terrell (1975) have applied both univariate and multivariate time series techniques to weekly-average, daily-maximum oxidant data. The results pertained to long-term prediction, using previous years' oxidant data, and could be improved by using oxidant data up to the week to be predicted. Little predictive information was found in week-old meteorological data, as might be expected from weather forecasting experience (Altshuller, 1975).

Linear multiple regression involves finding the linear combination of predictor variables, x_i, which best forecasts oxidant levels,

$$OX = \alpha_0 + \Sigma_i \alpha_i X_i + \varepsilon \quad ,$$

in which α_0, α_1,... are constants to be discovered and ε is an error term.

Chock, Levitt and Terrell (1975) have applied multiple regression to predict weekly averages of oxidant daily maxima from concurrent weekly average weather parameters, first using a regression analysis to screen out the best subset of independent predictors. Tiao, Phadke and Box, (1975) used a logarithmic regression model on data from Los Angeles to derive a forecasting relationship for daily maximum hourly oxidant based on the previous day's oxidant value, the month of the year, 4 A.M. NO_2 level, 4 A.M. inversion base height and its square, the difference between the inversion breaking temperature and the 4 A.M. surface temperature, and the average 1-4 A.M. wind speed.

Others using regression models include Breiman and Meisel (1976) and Bruntz, Cleveland, Kleiner, and Warner (1974).

There is no a priori reason to suppose that the relationship of oxidant to meteorological and pollutant predictors should be best fitted by any particular mathematical form. Methods employing formal pattern recognition have been studied, but to-date, no major predictive effort has been developed. In order to accomplish some of the objectives, two basic ideas are involved. First, out of the very large set of possible predictors, a smaller number of significant features must be selected, the aim being to find those which singly or in combination can best be used to forecast future oxidant levels. The next task is to find an optimal forecasting method linking these features to oxidant values. Multivariate piecewise linear regression could be used to approximate global nonlinearities in the "real" feature-oxidant relationship, yet still give continuous predictions. If prediction into categories is desired--for example, whether a given day will or will not exceed a particular oxidant standard--then many pattern classification methods are available.

McCutchan and Schroeder (1973), used discriminant analysis (a linear pattern recognition technique) to classify meteorological patterns in Southern California with high accuracy. They noted that certain patterns corresponded to high oxidant values.

An example of an approach that is in the spirit of pattern recognition, but does not use its formal mathematical methodology, is the objective ozone forecast system developed by Davidson (1974) at the Los Angeles Air Pollution Control District (LAAPCD) to predict the occurrence of days from July through October with ozone levels equal to or greater than 0.35 ppm. Details of this procedure are given in the Phase I report.

Real-time sulfate prediction has been pioneered in California and more specifically in the SCAB. Prediction methods have been restricted to either regression analysis or meteorological pattern recognition. In general, sulfate forecasting has been performed by the ARB and the AQMD through the prediction methods described by the California ARB (1976) and Zeldin et al. (1976). Subsequent models, while not specifically designed for prediction, have related sulfate levels to meteorological, air quality or emissions factors. Such work has been done by Cass (1975, 1976), White (1977), and Environmental Research and Technology (1977).

A review of recent literature failed to yield any meaningful information concerning ambient SO_2 prediction. Most efforts pertain to SO_2 point source dispersion models, rather than area-wide conditions. Three examples of current SO_2 models worth noting are by Shir and Shieh (1973), Goumans and Clarenburg (1975), and Gibson and Peters (1977). Each model attempted simulation of SO_2 concentrations for specified areas based upon emissions data and meteorological data. Although these models are not applicable to daily forecast procedures, they do express the progress that has been made in SO_2 modeling capabilities.

2.2 REVIEW OF EXISTING PREDICTION METHODS

The prediction of oxidants in the SCAB is routinely performed by the ARB and the South Coast Air Quality Management District (SCAQMD). (The National Weather Service [NWS] routinely predicts pollution potential based on meteorological criteria.) Basic techniques employ the use of regression analysis, point classification systems and air stagnation advisories. This section presents a limited summary of the existing predictive capabilities for oxidants as well as existing sulfate and SO_2 forecast algorithms.

ARB Procedures: Oxidants

Daily oxidant forecasts for each of the 19 statewide locations currently active on the ARB telemetry system are made by the ARB meteorology section. Predictions are initially issued at approximately 2 P.M. to be valid for the following day (Kinney, 1977). The initial prediction is updated on the morning of the valid date to add potential prediction resolution. One of the most significant oxidant forecasts for the Los Angeles air basin is Upland's one-hour max concentration.

A prediction equation for the Upland maximum hourly average is used to objectively determine the predicted concentration. Based on multiple regression techniques, the equation includes both meteorological and air quality data:

$$OX = 0.36A + 0.50B - 0.59C + 0.51D + 0.53E + 9.28 \qquad (2\text{-}1)$$

where

OX = predicted Upland maximum hourly average for tomorrow

A = today's Upland maximum hourly average between 0600-1400 PST

B = San Diego 500 mb height change 12Z* today - 12Z yesterday
C = temperature of LAX Inversion top from morning RAOB** (13Z)
D = temperature of 850 mb level at LAX from afternoon RAOB (19Z)
E = temperature of LAX Inversion top from afternoon RAOB (19Z)

In the sense that all input into the equation is based on in-hand data, the method is completely objective. The meteorologist, however, still may subjectively change the prediction derived from the equation, if conditions warrant (Kinney, 1974; CARB, 1975). The values for the remainder of the stations are entirely subjectively determined. However, since, in most cases, the Upland concentration represents the expected basin-wide maximum, the need for objective prediction guidance is not as critical for the other locations.

A same-day update equation is used again for the Upland station. Two distinct equations have been established: one for weekdays (Tuesday through Saturday), and another for Sunday and Monday. The primary difference between the two is that Equation (2-2) relies more heavily on the NO_2 concentrations, whereas Equation (2-3) depends more heavily on meteorological parameters:

Sunday-Monday

$$OX = 0.44A + 0.66B + 0.72C + 0.34D + 1.4 \qquad (2\text{-}2)$$

Tuesday-Saturday

$$OX = 0.42A + 1.04B + 0.21C + 0.75D - 1.4 \qquad (2\text{-}3)$$

where

OX = predicted Upland maximum hourly average for that day
A = Upland's maximum hourly average on the previous day
B = San Diego 500 mb height change 12Z today-12Z yesterday
C = downtown L.A. 0600 - 0900 PST NO_2 maximum hourly average
D = temperature of 850 mb level at LAX from morning RAOB (13Z)

* "Z" refers to Greenwich Meantime (GMT), which is 8 hours earlier than Pacific Standard Time (PST).

** RAOB is an abbreviation for Rawinsonde Observation, which is a low-level vertical atomspheric sounding.

It should be noted that, in developing these equations, high oxidant day data were input twice to force the equations to be more responsive to high oxidant prediction days. As a result, low-to-moderate oxidant days would tend to be overpredicted.

As in the case of the one-day prediction, results from the equations are used as a guide toward the issuance of a subjectively produced prediction, issued at approximately 10 A.M.

ARB Procedures: Sulfates

Sulfate prediction, which was initiated in June 1976 due to the ARB's promulgation of sulfate-oxidant episode criteria on May 28, 1976, was developed initially for Temple City (CARB, 1976). Using a screening technique, the ARB was able to filter out a majority of days in which meteorological conditions were not conducive to increased sulfate concentrations. The filter was defined as:

$SO_4^{=} < 20\ \mu g/m^3$ if

(1) 850 mb temperature anomaly <0, or

(2) Inversion base height <700 feet or >3300 ft, or

(3) Inversion magnitude (ΔT) < -7°C

From the remaining data, multiple stepwise regression techniques were employed to derive an equation predicting 24-hour sulfate concentrations:

$$S_{TEM} = \frac{2.3(A)^{.52}(D)^{.18}(E)^{.25}}{(B)^{.20}(C)^{.15}} \qquad (2\text{-}4)$$

where S_{TEM} = Sulfate concentration at Temple City ($\mu g/m^3$)

A = previous day's sulfate concentration at Temple City ($\mu g/m^3$)

B = EMT visibility at 23Z (miles)
C = LAX visibility at 23Z (miles)
D = LAX 20Z inversion base height
(hundreds of feet)
and E = LAX 20Z 1000 mb dewpoint (°C)

The results of this work indicate that high sulfate concentrations are related to low visibility, moderately deep and strong inversions, and the availability of moisture. The effect of persistence is quite noticeable, indicative of sequential sulfate build-up.

With the 1977 expansion of daily sulfate sampling to six SCAB sites (Temple City, Azusa, Anaheim, Reseda, Upland, and Riverside), the ARB pursued additional prediction equations for those locations. And with a need for earlier decision-making, visibility input data were changed from 23Z to 22Z. For each of the six generated equations, a screening process was employed, in the same manner as the 1976 effort. Results of the prediction filter are listed in Table 2.1.

From the remaining data, multiple stepwise regression yielded the following equations:

$$S_{TEM} = e^{[.49\ \ell nA_1\ -\ .18\ \ell nB\ -\ .18\ \ell nC\ +\ .22\ \ell nD_1\ +\ .23\ \ell nE_1\ +\ 0.84]} \tag{2-5}$$

$$S_{AZU} = e^{[.49\ \ell nA_2\ -\ .18\ \ell nB\ -\ .18\ \ell nC\ +\ .22\ \ell nD_1\ +\ .23\ \ell nE_1\ +\ 0.84]} \tag{2-6}$$

$$S_{ANA} = e^{[.23\ \ell nA_3\ -\ .21\ \ell nB\ -\ .32\ \ell nC\ -\ .19\ \ell nD_2\ +\ .06\ \ell nE_2\ +\ 3.10]} \tag{2-7}$$

$$S_{RES} = e^{[.45\ \ell nA_4\ -\ .08\ \ell nB\ -\ .17\ \ell nC\ -\ .50\ \ell nD_1\ +\ .57\ \ell nE_3\ -\ 1.27]} \tag{2-8}$$

$$S_{UPL} = e^{[.10\ \ell nA_1\ -\ .48\ \ell nB\ -\ .23\ \ell nC\ +\ .29\ \ell nA_6\ +\ .46\ \ell nE_3\ +\ 2.13]} \tag{2-9}$$

$$S_{RIR} = e^{[.47\ \ell n A_5\ -\ .17\ \ell n B\ -\ .16\ \ell n C\ +\ .27\ \ell n D_1\ +\ .03\ \ell n E_3\ +\ 1.08]} \qquad (4\text{-}10)$$

where S_{TEM} = Sulfate concentration at Temple City ($\mu g/m^3$)
S_{AZU} = Sulfate concentration at Azusa "
S_{ANA} = Sulfate concentration at Anaheim "
S_{RES} = Sulfate concentration at Reseda "
S_{UPL} = Sulfate concentration at Upland "
S_{RIR} = Sulfate concentration at Riverside "
A_1 = Previous day's sulfate concentration at Temple City
A_2 = Previous day's sulfate concentration at Azusa
A_3 = Previous day's sulfate concentration at Anaheim
A_4 = Previous day's sulfate concentration at Reseda
A_5 = Previous day's sulfate concentration at Upland
A_6 = Previous day's sulfate concentration at Riverside
B = EMT 22Z visibility
C = LAX 22Z visibility
D_1 = LAX 20Z inversion base (hundreds of feet)
D_2 = LOS 22Z SO_2 concentration
E_1 = LAX 20Z 1000 mb dewpoint
E_2 = LAX 20Z 850 mb temperature anomaly
E_3 = LAX 20Z 850 mb temperature
E_4 = ΔP(LAX-TRM), 22Z

These equations are currently in use by the ARB meteorology section. Predictions are issued by 3:00 P.M.

Table 2.1 Sulfate Prediction Filter

Sulfate predicted <2 μg/m^3 if <u>any</u> of the following conditions exist.

Location	Criteria
Temple City Azusa Anaheim Reseda	(1) LAX 20Z Inversion base <700 feet (2) LAX 20Z Inversion base >3500 feet (3) LAX 20Z Inversion magnitude (ΔT) <7°C (5°C in Winter, Azusa all year) (4) LAX 20Z 850 mb Temperature Anomaly <0°C (5) ΔP(LAX-TRM), 22Z <0 mb
Riverside Upland	(1) LAX 20Z Inversion base <700 feet (2) LAX 20Z Inversion base >4500 feet (3) LAX 20Z Inversion magnitude (ΔT) <4°C (4) ΔP(LAX-TRM), 22Z <0 mb

SCAQMD Procedures: Oxidants

Starting in April 1976, all basin-wide oxidant forecasts were issued as one forecast by the combined four-county meteorological staff of the SCAQMD centralized at El Monte. Daily sulfate forecasts were initiated in May 1977 following the adoption of a revised Regulation VII incorporating sulfates into the list of episode criteria.

The AQMD provides pollution forecasts 7 days a week during the smog season (approximately May through October) and 6 days per week during the remaining months. In reality, there is a forecast issued for every day of the year, however, during the non-smog season, the off day is covered by a 2-day prediction issued on the preceding work day (Keith, 1977).

For each daily prediction, the forecaster subjectively predicts the expected LAX inversion conditions for the following morning. Based on this prediction, three elements are completed:

(1) An objective forecast prediction model (Figure 2.1) for the San Bernardino maximum oxidant potential (based on a point classification system)

(2) An objective forecast prediction model (Figure 2.1) Los Angeles County maximum oxidant potential (based on multiple linear regression)

(3) A daily computer forecast worksheet

Using one of the in-house mini-computers, a pre-programmed forecast package is used. The forecaster inputs data from the daily computer forecast worksheet via teletype and the computer determines the entire forecast values by station. After examining the output, the forecaster can program any element change which he subjectively determines. The computer outputs the resultant forecast in the exact format necessary for dissemination (prior to 11 A.M.), including source-receptor locations, if applicable. An example of final computer output is presented in Figure 2.2.

DAILY AIR POLLUTION FORECAST

FORECAST issued_______________By_______________For_______________
time date forecaster day date

	BASE	TOP	MAG	850-SFC	950mb	24 HR SUM	15Z SUM
INV	ft	ft	°C	°C	°C	mb	mb
	STABILITY	950mb	INV	GRADIENT	DAY	MONTH	TOTAL
EQ							

VERIFICATION

	BASE	TOP	SFC	950mb	850 mb	INV BASE	INV TOP
LAX	ft	ft	°C	°C	°C	°C	°C
EMT	ft	ft	°C	°C	°C	°C	°C
	MAG		850-SFC			15Z GRAD	
	STABILITY	950	INV	GRAD	DAY	MONTH	TOTAL
EQ							

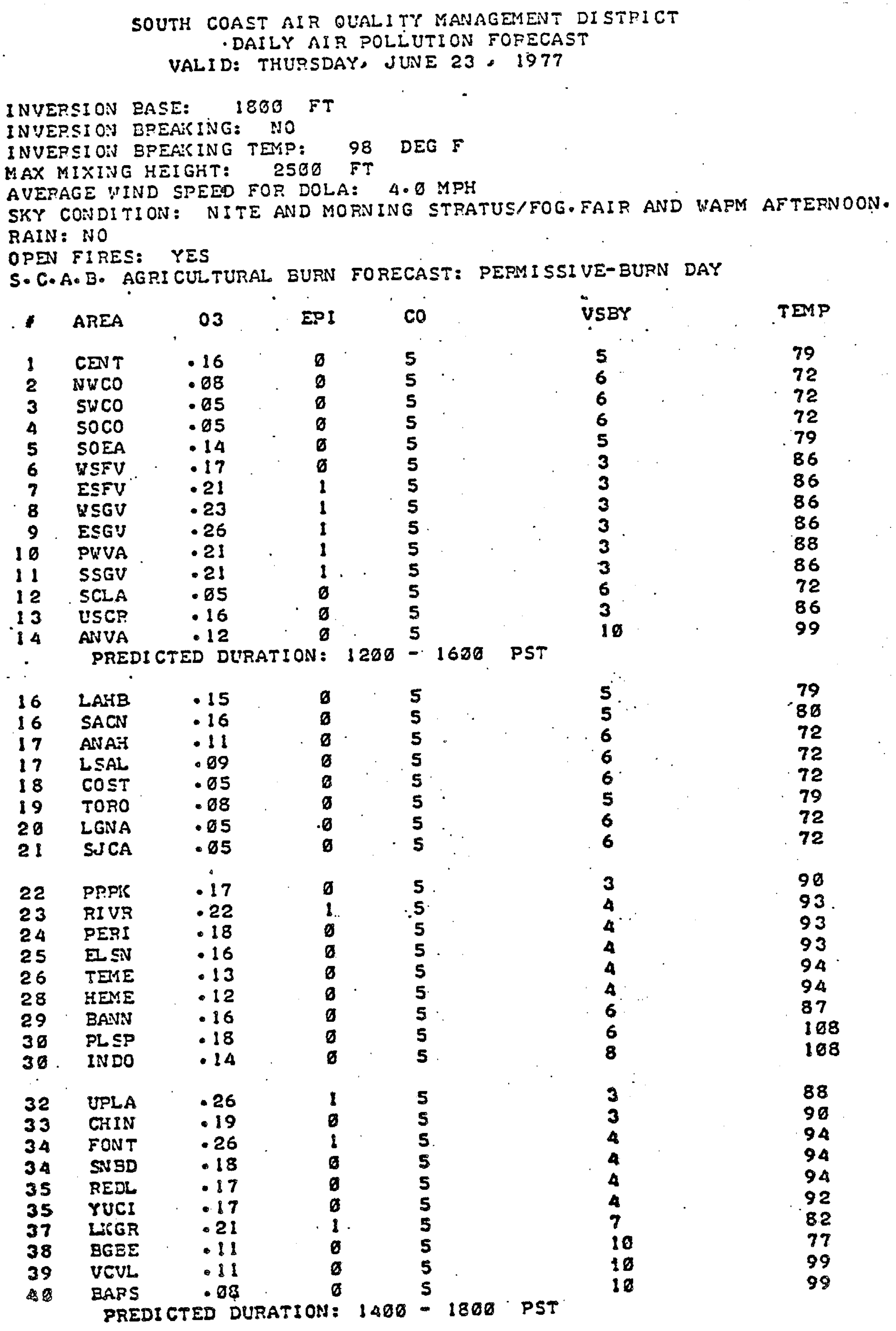
I-Ming FORECAST Model

VARIABLE		CONSTANT		MONTHLY ADJUSTMENT VALUES
$(T850)^2$		x .0193		Jan -5.5 Jul 0.2
T900		x .456		Feb -2.4 Aug 0.0
T950		x .305		Mar 0.8 Sep 0.7
T1000		x - .639		Apr 1.9 Oct -0.9
Monthly Adj		x 1.0		May 0.2 Nov -4.2
Factor	8.8	x 1.0	8.8	Jun 2.1 Dec -5.4
TOTAL			PPHM	

VERIFICATION Model

VARIABLE		CONSTANT		MONTHLY ADJUSTMENT VALUES
$(T850)^2$		x .0142		Jan -6.9 Jul 0.5
T1000		x - .303		Feb -4.3 Aug 0.0
Max OX LAS		x .29		Mar -0.6 Sep 0.3
H500, VBG		x .0194		Apr 1.2 Oct -2.8
dP LAX-DAG		x - .634		May 1.0 Nov -5.8
dP SAN-LAS		x -1.096		Jun 2.5 Dec -7.1
dP LAX-LAS		x .595		
dP LAX-WJF		x - .652		
Monthly Adj		x 1.0		
Factor	-97.1	x 1.0	-97.1	
TOTAL			PPHM	

Figure 2.1 Example of AQMD Objective Prediction Models



```
                    SOUTH COAST AIR QUALITY MANAGEMENT DISTRICT
                          ·DAILY AIR POLLUTION FORECAST
                        VALID: THURSDAY, JUNE 23 , 1977

INVERSION BASE:    18ØØ  FT
INVERSION BREAKING:  NO
INVERSION BREAKING TEMP:    98  DEG F
MAX MIXING HEIGHT:    25ØØ  FT
AVERAGE WIND SPEED FOR DOLA:   4.Ø MPH
SKY CONDITION:  NITE AND MORNING STRATUS/FOG.FAIR AND WARM AFTERNOON.
RAIN: NO
OPEN FIRES:  YES
S.C.A.B. AGRICULTURAL BURN FORECAST: PERMISSIVE-BURN DAY

 #   AREA      O3      EPI     CO        VSBY        TEMP

 1   CENT     .16       Ø      5          5           79
 2   NWCO     .Ø8       Ø      5          6           72
 3   SWCO     .Ø5       Ø      5          6           72
 4   SOCO     .Ø5       Ø      5          6           72
 5   SOEA     .14       Ø      5          5           79
 6   WSFV     .17       Ø      5          3           86
 7   ESFV     .21       1      5          3           86
 8   WSGV     .23       1      5          3           86
 9   ESGV     .26       1      5          3           86
1Ø   PWVA     .21       1      5          3           88
11   SSGV     .21       1      5          3           86
12   SCLA     .Ø5       Ø      5          6           72
13   USCR     .16       Ø      5          3           86
14   ANVA     .12       Ø      5          1Ø          99
        PREDICTED DURATION:  12ØØ - 16ØØ   PST

16   LAHB     .15       Ø      5          5           79
16   SACN     .16       Ø      5          5           8Ø
17   ANAH     .11       Ø      5          6           72
17   LSAL     .Ø9       Ø      5          6           72
18   COST     .Ø5       Ø      5          6           72
19   TORO     .Ø8       Ø      5          5           79
2Ø   LGNA     .Ø5       Ø      5          6           72
21   SJCA     .Ø5       Ø      5          6           72

22   PRPK     .17       Ø      5          3           9Ø
23   RIVR     .22       1      5          4           93
24   PERI     .18       Ø      5          4           93
25   ELSN     .16       Ø      5          4           93
26   TEME     .13       Ø      5          4           94
28   HEME     .12       Ø      5          4           94
29   BANN     .16       Ø      5          6           87
3Ø   PLSP     .18       Ø      5          6           1Ø8
3Ø   INDO     .14       Ø      5          8           1Ø8

32   UPLA     .26       1      5          3           88
33   CHIN     .19       Ø      5          3           9Ø
34   FONT     .26       1      5          4           94
34   SNBD     .18       Ø      5          4           94
35   REDL     .17       Ø      5          4           94
35   YUCI     .17       Ø      5          4           92
37   LKGR     .21       1      5          7           82
38   BGBE     .11       Ø      5          1Ø          77
39   VCVL     .11       Ø      5          1Ø          99
4Ø   BARS     .Ø8       Ø      5          1Ø          99
        PREDICTED DURATION:  14ØØ - 18ØØ   PST
```

Figure 2.2 Example of AQMD Computer Forecast Final Output

Prediction Techniques

The AQMD has three oxidant prediction aids, as mentioned previously. The method for each will be discussed here.

Objective System for San Bernardino

Ozone prediction for the eastern valley areas of the South Coast Air Basin (east of Pomona) is based on the premise that primary ozone precursors originate in the populous Los Angeles-Orange County metropolitan areas in the morning hours and are then transported northward and eastward. The photochemical processes continue during the daytime seabreeze transport, reaching the eastern Basin areas with peak ozone values late in the afternoon. Therefore, the ozone prediction model examines the meteorological potential for the buildup of contaminants in the morning plus the potential for transport during the day.

A point classification system is used to relate the given meteorological parameters to the expected peak ozone for San Bernardino. Points are assigned on a 0 to 10 scale with the more adverse conditions at the high end (see Table 2.2). The classification categories are defined as follows:

(1) Stability (°C) = $(T_{850mb} - T_{sfc}) + (T_t - T_b)$

where T_t = temperature of the inversion top, and T_b = temperature of the inversion case.

(2) 950 mb Temperature (°C)

(3) Inversion Base Height (Ft MSL)

(4) Gradient (mb) = $(P_1-P_2) + (P_3-P_4) + (P_5-P_6)$

where

P_1 = Long Beach (LGB) sea level pressure
P_2 = Daggett (DAG) sea level pressure
P_3 = San Diego (SAN) sea level pressure
P_4 = Las Vegas (LAS) sea level pressure
P_5 = Norton AFB (SBD) sea level pressure
and P_6 = George AFB (VCV) sea level pressure

(5) Day of the week

(6) Month of the year

Table 2.2 Point Classification System for Ozone Prediction Model

Points	Stability	950 Temp.	Inversion	Gradient	Day
0	≤ 5.0	≤ 5.0	5,000+	≤ -10.0 ≥ +20.0	
1	5.1 - 7.0	5.1 - 8.0	4,001 - 5,000	- 9.0 to - 9.9 +16.0 to +19.9	Sunday
2	7.1 - 9.0	8.1 - 11.0	3,001 - 4,000	- 8.0 to - 8.9 +12.0 to +15.9	
3	9.1 - 11.0	11.1 - 14.0	2,501 - 3,000	- 7.0 to - 7.9 +10.0 to +11.9	Monday Tuesday Wednesday Saturday
4	11.1 - 13.0	14.1 - 17.0	2,001 - 2,500 Surface	- 6.0 to - 6.9 + 8.0 to + 9.9	Thursday
5	13.1 - 15.0	17.1 - 20.0	1,501 - 2,000	- 5.0 to - 5.9 + 6.0 to + 7.9	
6	15.1 - 17.0	20.1 - 24.0	1,001 - 1,500	- 4.0 to - 4.9 + 4.0 to + 5.9	Friday
7	17.1 - 19.0	24.1 - 28.0	701 - 1,000	+ 2.0 to + 3.9	
8	19.1 - 21.0	28.1 - 32.0	501 - 700	- 3.0 to - 3.9 + 1.0 to + 1.9	
9	21.1 - 23.0	32.1 - 36.0	301 - 500	- 2.0 to - 2.9 0.0 to + 0.9	
10	≥ 23.1	≥ 36.1	150 - 300	- 1.9 to - 0.1	

Month	
Jan.	-15
Feb.	-12
Mar.	- 9
Apr.	- 5
May	0
June	+ 1
July	+ 2
Aug.	+ 2
Sep.	+ 1
Oct.	8
Nov.	-11
Dec.	-15

The first three categories are based on data obtained from the morning Los Angeles International Airport (LAX) sounding, taken daily at approximately 1430 GMT. The pressure gradients, (4), are obtained from the 1500 GMT surface observations. In addition, two categories (day of the week and month of the year) complete the model for additive equivalency to the expected maximum hourly average.

This system is used in two ways. Predicted inversion and pressure gradient data are used to compute a value for the following day. In a sense, this results in subjective information inserted into an objective model. Results, therefore can only be expected to be as good as the subjective capability of the forecaster. In the second approach, same-day data are used to produce a completely objective same-day prediction (with approximately 6 to 12 hours lead-time).

Objective System for Los Angeles County

Using multiple regression techniques, a model was developed for predicting the daily maximum hourly average in Los Angeles County. The model was generated using 1974-1975 meteorological and oxidant data. Preliminary results yielded two equations:

Day-in-advance:

$$OX = -0.67\ T_1 + 0.36\ T_2 + 0.35\ T_3 + 0.23\ T_4 + 0.35\ T_5 + D + M + 6.2$$

Same day:

$$OX = -0.36\ T_1 + 0.66\ T_3 + 0.25\ X_1 - 0.56\ P_1 - 0.81\ P_2 + D + M + 11.2$$

where:

T_1 = 1000 millibar (mb) temperature (°C)
T_2 = 900 mb temperature (°C)
T_3 = 850 mb temperature (°C)
T_4 = Inversion base temperature (°C)
T_5 = Inversion top temperature (°C)
X_1 = Maximum hourly OX average on the previous day (PPHM)
P_1 = Los Angeles to Palmdale pressure gradient (mb)
P_2 = Long Beach to Daggett pressure gradient (mb)
D = Constant term for day-of-the-week
and M = Constant term for month-of-the-year

It should be noted that the day-in-advance equation requires subjectively derived input. As in the objective system described in the previous subsection, this approach is not a truly objective prediction method. The same-day equation, however, does use only available data, and is therefore completely objective.

Results from these equations, when used on 1976 data, showed weaknesses in predicting oxidant concentrations ≥ .30 ppm. In fact, the model never predicted above .30 ppm. Therefore, the set of equations was revised using 1973-1976 data as the dependent data set (since both 1973 and 1976 contained a greater number of days ≥ .30 ppm than 1974 and 1975). Results of multiple linear regression yielded a new set of equations, which is given in the lower half of Figure 2.1.

Computer Prediction Model

Similar to the objective methods previously described, the AQMD computer prediction model (CPM) is based upon subjectively determined meteorological information. However, unlike the other methods, the CPM is able to detect meteorological inferences which can affect not only the concentration levels but also the distribution pattern. For example, both the San Bernardino and Los Angeles models yield a numerical value predicting a maximum hourly average. Equivalent model values can be obtained under a variety of meteorological conditions. The CPM, while using both models as a guide to concentration levels, is able to adjust the distribution pattern based on the meteorological differences.

Also, the logic of the CPM was developed to improve the known weaknessess of the objective models. Under certain meteorological conditions, the San Bernardino model is better than the Los Angeles model, and at other times, the opposite is true. The CPM essentially weights the prediction according to the method which is most favorable. The CPM tailors the oxidant forecast base upon the contributions of differing meteorological variables (for example inversion strength and the direction and strength of the pressure gradients).

It should be noted that the CPM has a manual override function in which the forecaster can subjectively change any one or more predicted values. Additionally, the forecaster can change the designated routing (i.e., northern route to southern route) and receive a re-computed forecast distribution. There are other forecast values for carbon monoxide, temperature, and visibility; however, the details of these features are beyond the scope of this study.

SCAQMD Procedures: Sulfates

Similar to the ARB, the AQMD developed a sulfate prediction capability in mid-1976. But unlike the ARB which used both air quality and meteorological data, the AQMD devised a two-dimensional nomogram based solely on inversion base height and strength (ΔT) (see Figure 2.3).

From this analysis, it can be seen that high sulfate days occur within a large range of inversion heights from 600 feet to near 4500 feet, the most critical area being between 1200 feet and 2400 feet, depending on the inversion intensity. Not unexpectedly, there is an extremely strong relationship between the nomogram and the ARB filtering criteria.

Using the nomogram, AQMD meteorologists use the subjectively determined inversion base height and magnitude (as input to the oxidant prediction) and determine a predicted sulfate value for the next day. Because the input does not require any air quality persistence data, such a prediction can be generated by 10 A.M. On the verifying day, actual inversion data are used in the nomogram to (1) verify the original meteorological prediction, and (2) update the sulfate forecast as necessary. Data for verification are routinely available by 9:00 A.M.

It should be noted that the nomogram does not represent a site-specific prediction. Rather, the prediction is indicative of a basin-wide maximum, regardless of location.

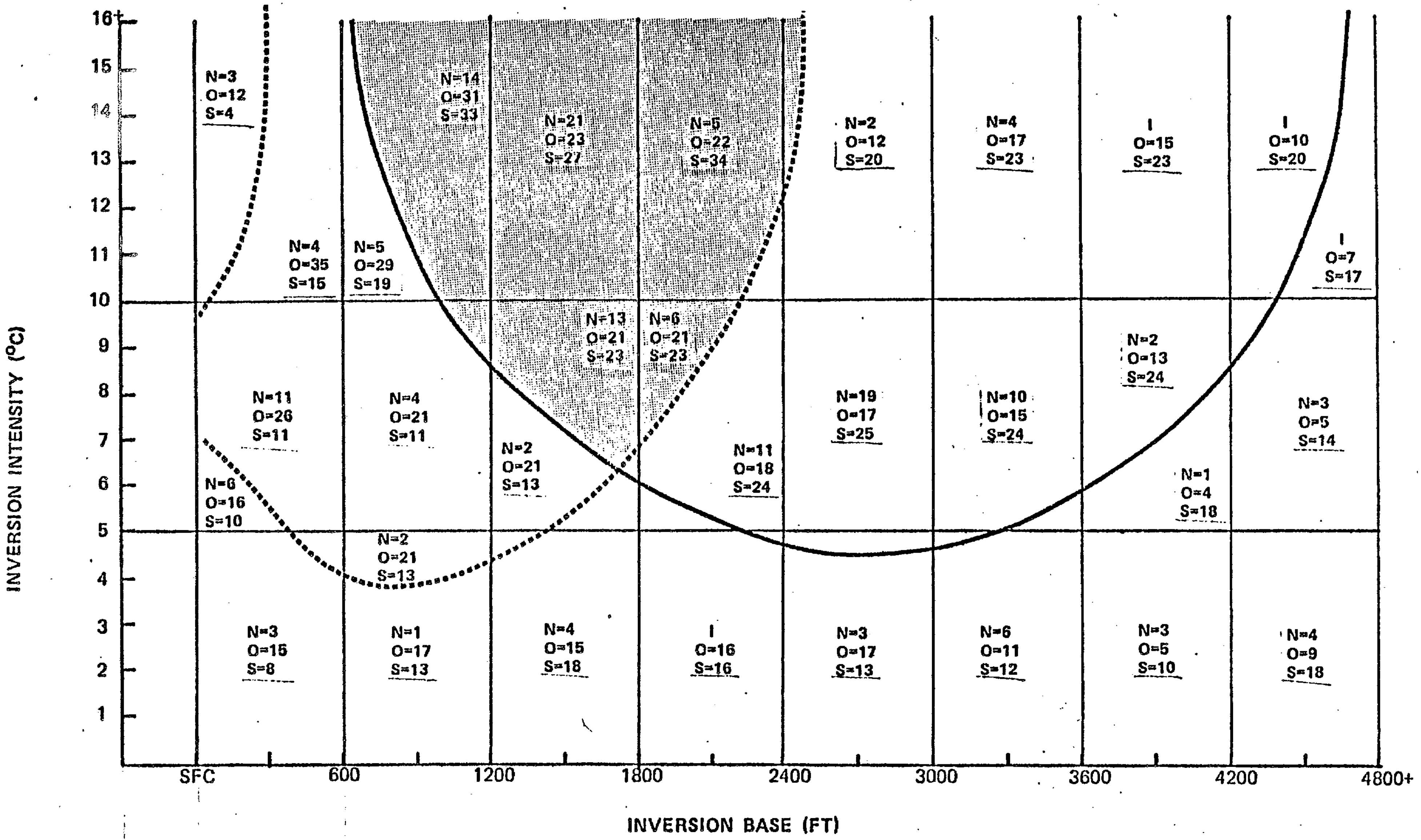


Figure 2.3 Distribution of ozone (pphm) at Azusa and sulfate ($\mu g/m^3$) at Glendora under specified inversion conditions. Areas of high sulfate values (solid curve) and high ozone (dashed curve) are given. For each interval, values represent average of observed data points where N=number of cases, O=observed ozone, S=observed sulfate, and I=interpolated values for ozone and sulfate where no data points exist. Shaded area is high sulfate and high ozone potential.

National Weather Service: All Pollutants

As of 1975 the Weather Service Forecast Office in Los Angeles (WSFO-LA) has been issuing two major statements of possible high pollution: Air Stagnation Advisories (ASA) and Special Dispersion Statements (SDS). ASA or SDS statements are issued when it is determined that meteorological conditions are conducive to the build-up of pollutant concentrations. An example of the criteria to determine the ASA-SDS is given in Table 2.3.

2.3 EXISTING VERIFICATION METHODS

Evaluation Methods--Oxidant

In establishing a baseline performance for existing oxidant prediction methods it is necessary to consider the various important features of prediction. Some pertinent questions considered were:

(1) How well do we predict episodes?

(2) How close do we come to predicting actual concentrations?

(3) How well are we able to predict those days in which significant changes occur (e.g., large deviations from persistence)?

(4) How well can we predict stage 2 episodes?

(5) What are the tradeoffs in making too many stage 2 predictions in order to catch all the episode days? (In other words, at what point does a high false alarm rate degrade the credibility of the product?)

To answer these questions, two primary analyses were conducted: (1) episode prediction analysis, and (2) quantitative prediction analysis.

To test episode level prediction accuracy, a "prediction contingency table" was constructed. A sample output is shown in Table 2.4. Number of occurrences for each predicted and observed episode level were tabulated. Correct predictions are indicated in the diagonal of the matrix

Table 2.3 ASA Guide and Warning List

1. Check NMC stagnation chart.

2. From LAX & EMT AM soundings, note existence of:

   a) Strong inversion, (8-10 degrees C or more).

   b) Steep slope, (15 MB or less from top to bottom of inversion) and mechanism for maintaining steep slope (usually cutoff high at 850 MB and 700 MB over Nevada with winds over Southern California NE to at times SE above the inversion).

   c) Base of inversion AM 1500 ft or lower and to be maintained near this height or lower during the day by continued subsidence aloft. High 500 MB heights with thermal trough in evidence but not necessarily well developed.

   d) Temperature greater than 25 degrees C at top of inversion (3500 ft or lower), summer months.

   e) LAX PM sounding, continued low inversion, 800 ft or less. EMT PM sounding low inversion and as low as or lower than AM sounding with mixing heights 1500 ft or less.

3. High point totals from objective systems.

   a) San Bernardino: 25-30 marginal, greater than 30, yes. (Uncorrected)

   b) Davidson: greater than .35 (may be low when SBD system is high and vice versa)

4. Subjective weight given to special factors as outlined in sections 4.2 and especially 5.1 of Chapter C-30, Air Pollution Weather Forecasts, Operations Manual.

Table 2.4 Tabular Layout of Episode Prediction Contingency Table

[illegible]REC[illegible]

OBSERVED

	LT 20	20 - 34	35 - 49	GE 50	TOTAL	POSSIBLE CORRECT	FALSE ALARM
LT 20	n_{00} n_{00}/N	n_{01} n_{01}/N	n_{02} n_{02}/N	n_{03} n_{03}/N	N_0 N_0/N	n_{00}/N_0	S_0 S_0/N_0
20 - 34	n_{10} n_{10}/N	n_{11} n_{11}/N	n_{12} n_{12}/N	n_{13} n_{13}/N	N_1 N_1/N	n_{11}/N_1	S_1 S_1/N_1
35 - 49	n_{20} n_{20}/N	n_{21} n_{21}/N	n_{22} n_{22}/N	n_{23} n_{23}/N	N_2 N_2/N	n_{22}/N_2	S_2 S_2/N_2
GE 50	n_{30} n_{30}/N	n_{31} n_{31}/N	n_{32} n_{32}/N	n_{33} n_{33}/N	N_3 N_3/N	n_{33}/N_3	S_3 S_3/N_3
TOTAL	M_0 M_0/N	M_1 M_1/N	M_2 $M_2/$	M_3 M_3/N	N 1.00	1.00	
CORRECT PREDICTNS	n_{00}/M_0	n_{11}/M_1	n_{22}/M_2	n_{33}/M_3	1.00		TOTAL $\sum_j S_j$ $\sum_j S_j/N$
FAIL TO ALARM	T_0 T_0/M_0	T_1 T_1/M_1	T_2 T_2/M_2	T_3 T_3/M_3		TOTAL $\sum_k T_k$ $\sum_k T_k/N$	

where: $N_j = \sum_{k=0}^{3} n_{jk}$; j=0-3 $\qquad$ $S_j = \sum_{k=j+1}^{3} n_{jk}$; j=0-2

$M_k = \sum_{j=0}^{3} n_{jk}$; k=0-3 $\qquad$ $T = \sum_{j=k+1}^{3} n_{jk}$; k=0-2 represented by equivalent episode predictions and observations. Also tabulated were the false-alarms, fail-to-alarms, and prediction accuracies Tabulated figures indicate both the number of occurrences (upper number) and the fractional amount of the total.

For an evaluation of concentration prediction accuracy, a tabular listing was devised by stratifying observed values according to increments of 5 pphm. For each designated class interval, the following parameters were computed:

(1) N_i: the number of observations per class interval

(2) the mean observed value

(3) the mean of the corresponding predicted values

(4) the mean absolute error:

$$\frac{\Sigma \, |\text{predicted value - observed value}|}{N_i}$$

(5) the number of predictions (and the fractional amount of N_i) that were ±2 pphm of the observed values. Similar tabulations made for ±5, ±8, ±10, ±15 pphm.

A sample of the layout is shown in Table 2.5.

Evaluation Criteria

Using the "raw" evaluation data including the episode prediction analysis and the quantitative prediction analysis, a more comprehensive evaluation criteria was determined. Analyses were first stratified by season: May through October, and November through April. This technique allowed for a meaningful assessment of smog-seasonal prediction without undue influence from the persistent low values occuring in the winter. On the other hand, the rare high-oxidant occurrences in winter can be separated from the routine summertime occurrences, such that specific analyses can be made for these off season events.

With the data separated into seasons a more definitive evaluation was performed using a series of individual criteria assessments. The significance of each individual criterion is given in the following brief summary. The cumulative result is a standardized evaluation methodology.

Table 2.5 Tabular Layout of Stratified Quantitative Error Analysis

RIVR MAXIMA 1977
MAY-OCT

TWO-DAY PERSISTENCE

INTERVAL FOR DAYD	N_i	MEAN OBSERVE VALUES	MEAN PREDICT VALUES	MEAN ABSOLUT ERROR	PRCNT WITHIN 2	PRCNT WITHIN 5	PRCNT WITHIN 8	PRCNT WITHIN 10	PRCNT WITHIN 15
LE 5	17.	3.7	7.6	4.2	7. 0.41	11. 0.65	15. 0.88	16. 0.94	17. 1.00
6 - 9	24.	7.8	11.5	5.3	10. 0.42	17. 0.71	19. 0.79	21. 0.88	22. 0.92
10 - 14	41.	11.8	14.9	6.2	8. 0.20	23. 0.56	30. 0.73	32. 0.78	39. 0.95
15 - 19	37.	16.9	18.0	5.7	6. 0.16	19. 0.51	33. 0.89	34. 0.92	37. 1.00
20 - 24	42.	22.0	17.5	6.5	8. 0.19	22. 0.52	29. 0.69	33. 0.79	40. 0.95
25 - 29	19.	26.6	22.6	4.8	6. 0.32	12. 0.63	17. 0.89	17. 0.89	19. 1.00
30 - 34	4.	30.8	18.8	12.0	0. 0.00	1. 0.25	1. 0.25	2. 0.50	3. 0.75
35 - 39	0.	0.0	0.0	0.0	0. 0.00	0. 0.00	0. 0.00	0. 0.00	0. 0.00
40 - 44	0.	0.0	0.0	0.0	0. 0.00	0. 0.00	0. 0.00	0. 0.00	0. 0.00
45 - 49	0.	0.0	0.0	0.0	0. 0.00	0. 0.00	0. 0.00	0. 0.00	0. 0.00
GE 50	0.	0.0	0.0	0.0	0. 0.00	0. 0.00	0. 0.00	0. 0.00	0. 0.00
TOTALS	184.	15.8	15.9	5.9	45. 0.24	105. 0.57	144. 0.78	155. 0.84	177. 0.96

Number of Occurrences (n_i)

$\frac{n_i}{N_i}$

1. Episode Prediction Accuracy

One method to evaluate prediction capability is to examine the ability to correctly predict episode levels regardless of the specific concentration (e.g., a prediction of 0.20 ppm is considered a stage 1 "hit" even if the observed value was 0.34 ppm, or a "miss" if the observed value was 0.10 ppm).

2. Quantitative Prediction Accuracy

While accurate episode prediction is a desirable goal, the ability to accurately predict pollutant concentration is a more formidable assessment of prediction capability. Several methods of analysis are possible: (1) compare the mean absolute errors of the different models; (2) examine the distribution of predicted values in light of the mean of the observed classes; (3) determine the percentage of predictions ±2 pphm of the observed values (a numerical list is defined as ±2 pphm of the observed value).

3. Significant Change Analysis

A key factor in determining the skill of a pollutant prediction method is to analyze the results of those days in which significant changes occurred from one day to the next (i.e., persistence is ineffective). To accomplish this, analyses were performed, to determine the number of predictions within ±2 pphm on days when significant changes of ≥ 10 pphm in in ozone concentrations occurred.

4. Stage-2 Episode Prediction Analysis

The ability to accurately predict stage-2 episode days is highly desirable. Obviously, it is advantageous to correctly predict all stage-2 days. Such days require significant abatement strategies to be implemented. On the other hand, too many false-alarm situations (i.e., causing abatement strategies to be implemented when in reality conditions did not prove to be that adverse) could degrade the credibility of the product, such that large-scale cooperation in implementing strategies could be severely limited. The best method would achieve an optimum condition which would gain the greatest public credibility and acceptance.

In regard to specific stage-2 oxidant prediction, one can argue that predictions can be made objectively--without the external influences caused by a stage-2 prediction. However, past experience has demonstrated a tendency to be cautious with respect to such predictions. Days on which a truly unbiased prediction would have resulted in a 35 pphm prediction, may in reality have been issued as 34 pphm. In other words, "if the condition is marginal, do not predict it."

Realizing, then, that the optimum situation is to keep the false-alarm rate at a minimum while attempting to increase the correct stage-2 predictions (i.e., increasing the probability of detection) a two-dimensional scoring system was developed to evaluate existing Stage-2 prediction capabilities. (See Figure 2.4). The system was weighted toward crediting those methods which kept a low false-alarm rate. Conversely, a high false-alarm rate was penalized more than a low probability of detection. It should be noted that several evaluation methods were attempted, including "skill score" based on the equation:

$$S = \frac{R - E}{T - E}$$

where

S = skill score

R = total number correct predictions

T = total number of predictions

E = expected number correct due to change

The major difficulty in using this method of evaluation was the indifference between false-alarm and fail-to-alarm situations. Because of the potential economic impact and loss of public credibility in false-alarm situations, it became necessary to develop an evaluation method which penalized excessive false alarms approximately twice that of fail-to-alarms. Figure 2.4 thus represents a heuristically determined scoring system designed to credit low false-alarm rates. Scoring values were assigned to the isopleths to assist in the numerical evaluation of prediction methods and to allow for reasonable improvement of effort in Phase II.

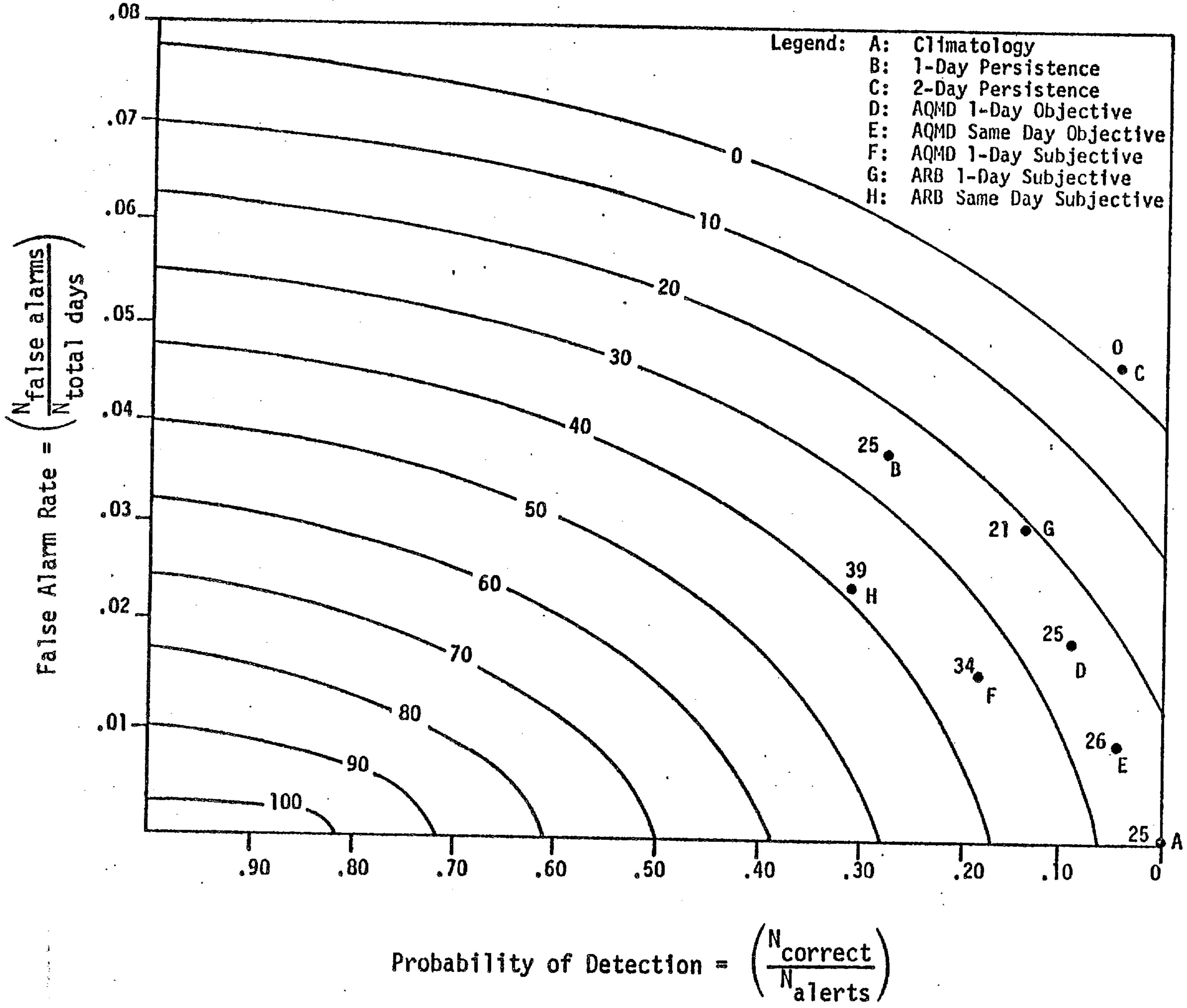


Figure 2.4 Scoring System Evaluation of Existing Stage 2 Episode Capabilities for Upland as a Function of the Probability-of-Dectection and False-Alarm Rate

5. Overall Baseline Prediction Evaluation

We have presented four significant considerations for evaluating prediction capabilities. To reflect a total prediction capability, an overall one-dimensional prediction rating method was devised using the equation:

$$\text{RATING} = T_C - 10E + T_2 + C + P$$

where:

T_C = the percentage of correct episode predictions

E = the mean absolute error

T_2 = the percent of predictions that were within ±2 pphm of observed values (an assessment of quantitative accuracy)

C = the percent of predictions that were within ±2 pphm of observed values on days in which observed values changed at least 10 pphm from one day to the next (an assessment of prediction capability in significant change situtations)

P = the Stage-2 prediction evaluation score from Figure 2.4

An example of an overall baseline evaluation is given in Table 2.6 From the numerical scores it can be seen that the best existing prediction method for Upland is the ARB same day subjective forecast.

2.4 EVALUATION METHODS--SULFATES

Similar methods used in the baseline performance evalution of oxidant prediction were employed to evaluate the sulfate prediction capability:

(1) the ability to predict episodes

(2) the ability to predict quantitative levels

(3) the tradeoffs between false-alarm and fail-to-alarm conditions.

(Significant change conditions will not be evaluated for sulfates because the number of samples in the generated subset is too small for valid conclusions.)

Table 2.6 Overall Prediction Rating for Upland--1974-1976 (May-Oct)

Number of Predictions		Method	T_c	-	10E	+	T_2	+	C	+	P	=	Overall Rating
		PERFECT	100	-	0	+	100	+	100	+	100	=	400
(460)	A.	Climatology	59	-	67	+	22	+	17	+	25	=	56
(454)	B.	1-Day Persistence	67	-	57	+	28	+	17	+	25	=	63
(452)	C.	2-Day Persistence	55	-	76	+	16	+	0	+	0	=	-5
(434)	D.	AQMD 1-Day Objective	68	-	56	+	30	+	14	+	25	=	81
(460)	E.	AQMD Same Day Objective	74	-	45	+	35	+	24	+	26	=	114
(460)	F.	AQMD 1-Day Subjective	68	-	57	+	28	+	12	+	34	=	85
(451)	G.	ARB 1-Day Subjective	68	-	54	+	29	+	17	+	21	=	81
(452)	H.	ARB Same Day Subjective	74	-	46	+	37	+	34	+	39	=	138

Since sulfate prediction began in mid-1976, only the June through October period of 1976 has been evaluated in this analysis. Data were evaluated using both the episode contingency tables and stratified quantitative error analysis formats as used in the oxidant evaluation. Verification intervals were changed in the contingency tables to reflect episode levels of ≥ 25 g/m^3.

Evaluation of Results

Incorporating the major aspects of prediction capabilities, a one-dimensional rating system was developed to compare the prediction methods. The overall rating was based on the equation:

$$\text{RATING} = T_c - 10E + T_2 + P$$

where

T_c = the percentage of correct episode predictions

E = the mean absolute error

T_2 = the percent of predictions that were within ±2 μg/m^3 of observed values

P = the episode prediction evaluation score from Figure 2.5.

Sample results are shown in Table 2.7.

It is worth noting that since the ARB equation utilizes an initial filter to screen out low potential sulfate days, it was not possible to evaluate the quantitative accuracy of the filtered days. Thus, the quantitative analysis includes only those days for which numerical values were predicted. (All days were included in the episode prediction evaluation, since filtered days represented non-episode forecasts). As a result, quantitative analysis produced a limited subset of days for which predicted values were made. In comparing E (the mean absolute error) and T_2 (% within ±2 μg/m^3), a disproportionate sample existed between persistence and the ARB equation. Persistence included all low days; the ARB equation included only high sulfate potential days. A quantitative comparison would therefore tend to bias the results in favor of persistence. To remedy the inconsistency, only those class intervals ≥ 15 μg/m^3 were used for the persistence computations of E and T_2. Using this procedure, the comparisons shown in Table 2.7 represent the best estimates between persistence and the ARB equations.

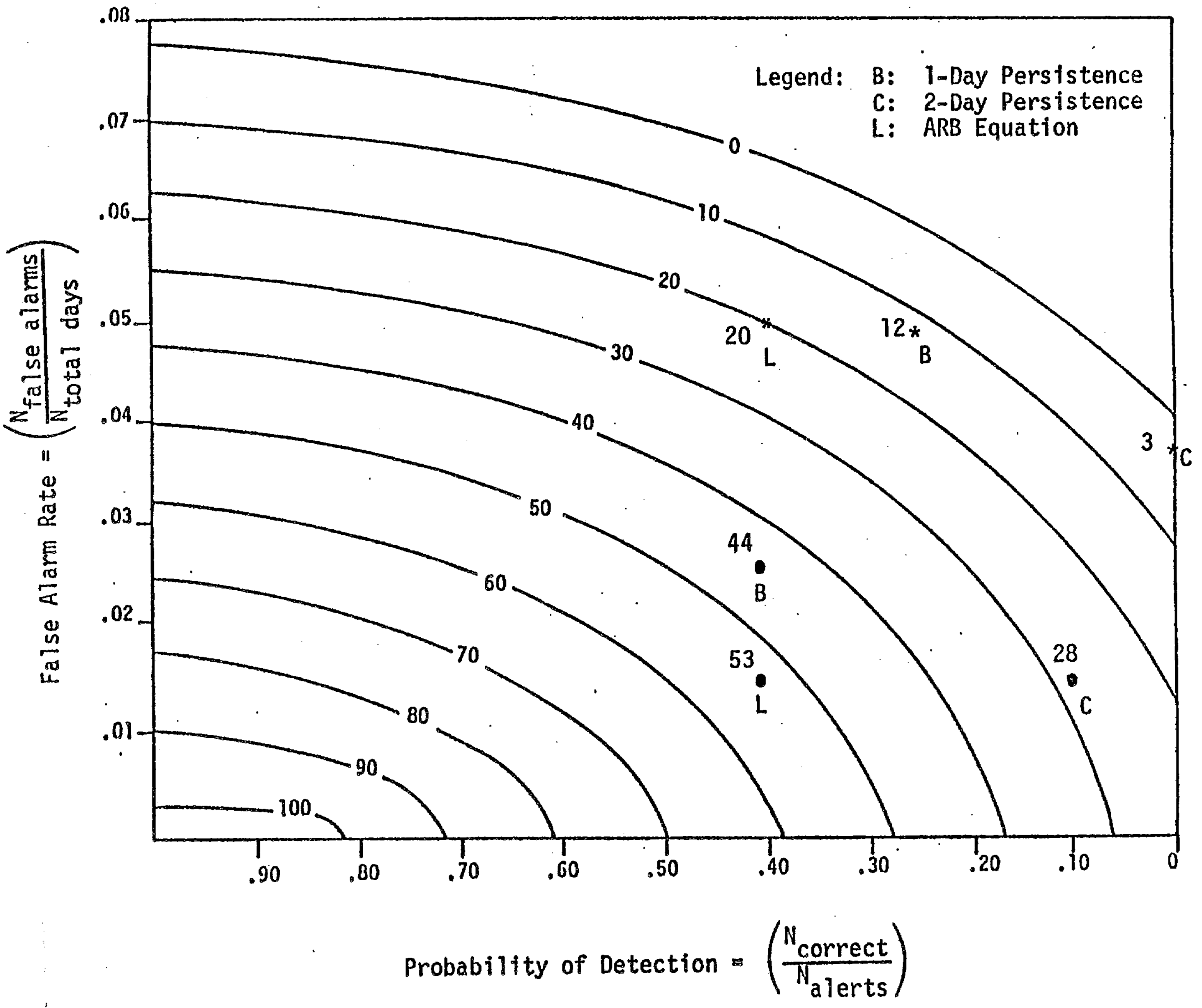


Figure 2.5 . Scoring system evaluation of existing sulfate episode prediction capabilities for Upland (*) and Temple City (●) as a function of the probability-of-detection and false-alarm rate.

Table 2.7 Overall Sulfate Prediction Rating (1976)

Number of Predictions	Method	T_c	- 10E	+ T_2	+ P	= Overall Rating
	Perfect	100	- 0	+ 100	+ 100	= 300
TEMPLE CITY						
	B.	94	- 50	+ 28	+ 44	= 126
	C.	92	- 60	+ 23	+ 3	= 58
	L.	95	- 34	+ 50	+ 53	= 164
UPLAND						
	B.	88	- 62	+ 15	+ 20	= 61
	C.	87	- 76	+ 18	+ 12	= 41
	L.	91	- 53	+ 26	+ 28	= 92

B = 1-Day Persistence

C = 2-Day Persistence

L = ARB Equation

Normalized Prediction Evaluation

The prediction rating procedures as given in Tables 2.6 and 2.7 allow for a quantitative one-dimensional evaluation of existing prediction capabilities. By normalizing with respect to the "perfect" prediction, we can compare all one-day-in-advance prediction methods to a baseline standard, e.g., 1-day persistence. In essence, we can measure the ability to beat 1-day persistence using the equation:

$$S = \frac{M-P_1}{PR}$$

where

S = score (comparable ability)

M = method rating (from Tables 2.6 or 2.7)

P_1 = 1-day persistence rating (from Tables 2.6 or 2.7)

PR = "Perfect" rating (from Tables 2.6 or 2.7)

An example of the normalized prediction evaluation is given by Table 2.8.

Table 2.8 Comparability of Day-in-Advance Prediction

Contaminant	Location	Method*	Perfect Rating (PR)	Method Rating (PR)	-	1-Day Persistence (PR)	=	Comparable Ability
OX	Upland	D	400	.202	-	.157	=	+.045
OX	Upland	F	400	.212	-	.156	=	+.055
OX	Upland	G	400	.202	-	.157	=	+.045
OX	DOLA	F	325	.335	-	.461	=	-.126
OX	DOLA	G	325	.455	-	.461	=	-.006
$SO_4^=$	Upland	L	300	.306	-	.203	=	+.103
$SO_4^=$	Temple City	L	300	.547	-	.420	=	+.127

*D = AQMD Obj.

F = AQMD Subj.

G = ARB Subj.

L = ARB Eq.

2.5 DATA BASE

Dependent Data Set 1974-1976

In order to develop pollution prediction algorithms a data base including approximately 400 air quality and meteorological variables for the 3-year period (1974-1976) was created. The air quality data base was comprised of ozone, SO_2, sulfate, NO_2, and total hydrocarbon data. (Records of air quality data prior to 1974 are of somewhat doubtful value for the purpose of short-term forecasting because of possible trend changes and in some cases changes in monitoring technology standards.) Meteorological variables included historical surface data and historical upper air data.

The bulk of the air quality-meteorological data base was assembled in a digitized form and input into the computer using a seven character alphanumeric code. The coding system was developed for convenience and flexibility allowing for easy manipulation of potential predictors. A complete list of all dependent data set variables is presented along with the descriptive code in Appendix A.

Air quality data from 39 stations (shown in Figures 2.6 - 2.8) and basin-wide pollution maxima in the South Coast Basin for differing pollutants were assembled. Included were selected averages of pollutant concentrations for the different time periods, 8-11 A.M., 6-9 A.M., 10 A.M. - 10 A.M. (24 hour average) and 12 A.M. - 12 A.M. (24 hour average), dependent upon the pollutant and the monitoring station. SO_2 emissions data from the Haynes and Los Alamitos power plants were also input into the data base.

Meteorological data for (1974 - 1976) were selected as a part of comprehensive network of variables directly and indirectly associated with local pollution phenomena. Historical surface data for 20 stations (shown in Figure 2.9) included such variables as wind direction and speed, surface temperature, visibility, pressure and temperature gradients (also 24 hour changes) and dewpoints. These variables

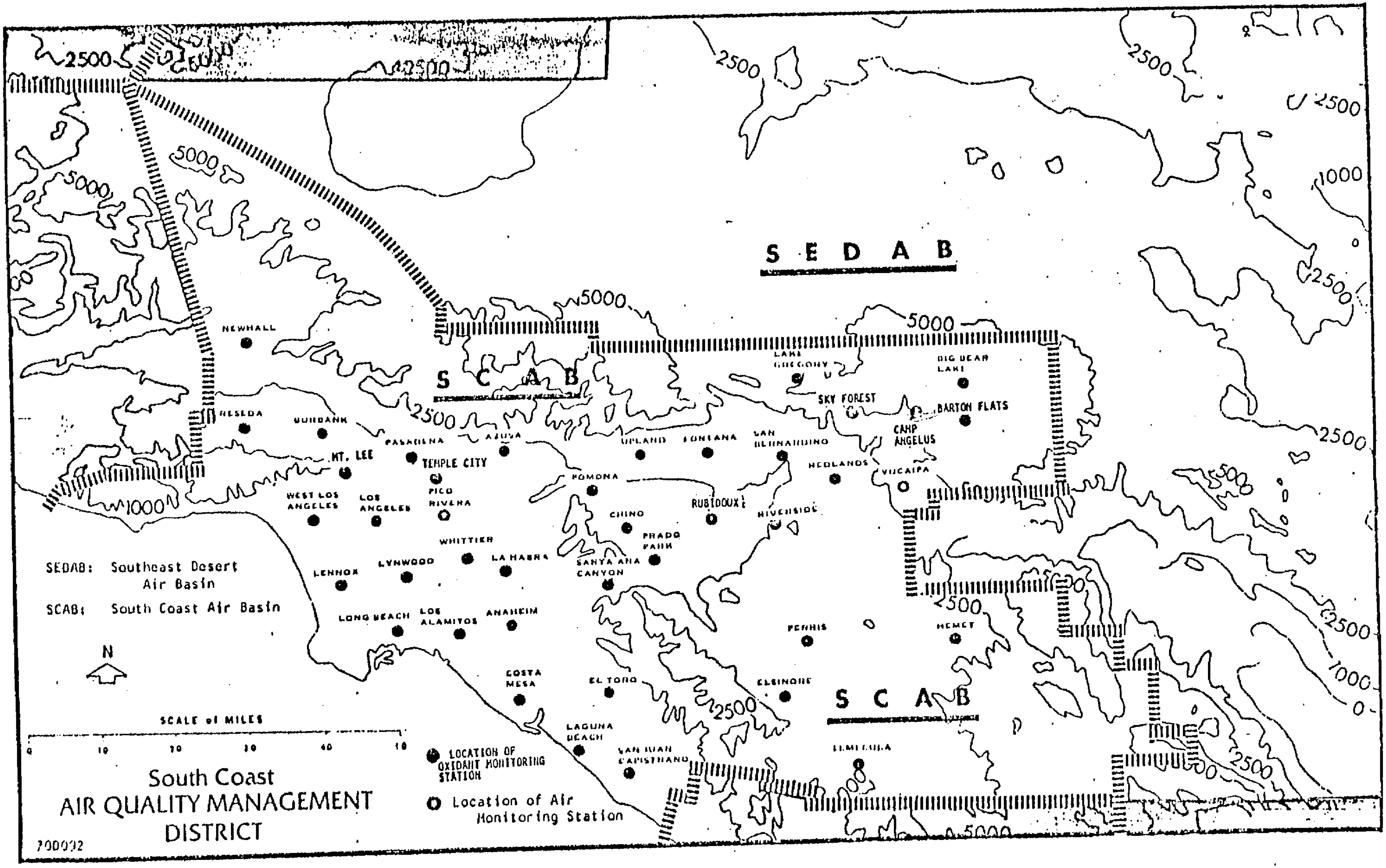


Figure 2.6 Location of SCAB Oxidant Monitoring Sites

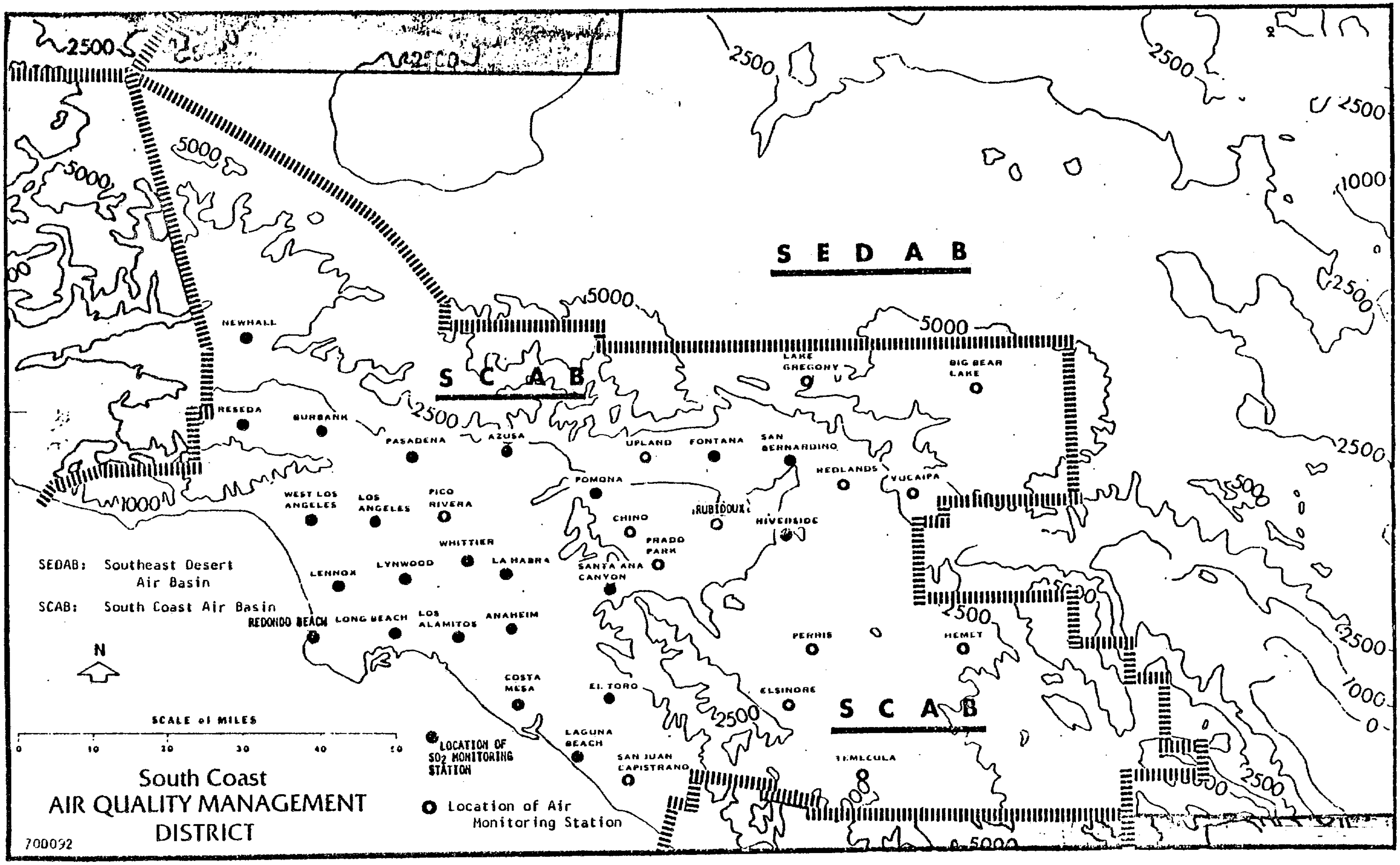


Figure 2.7 Location of SCAB Sulfur Dioxide Monitoring Sites

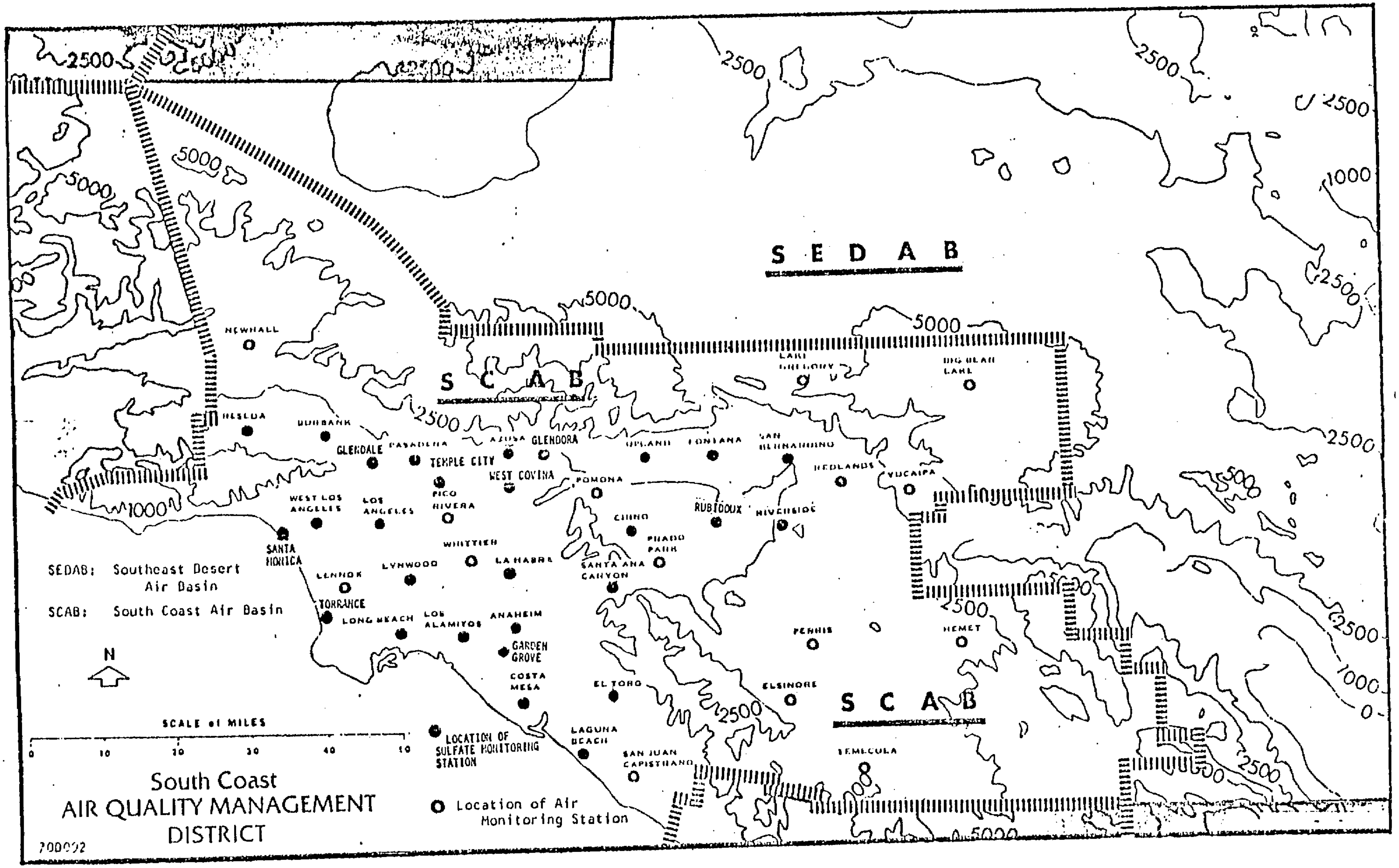


Figure 2.8 Location of SCAB Sulfate Monitoring Sites

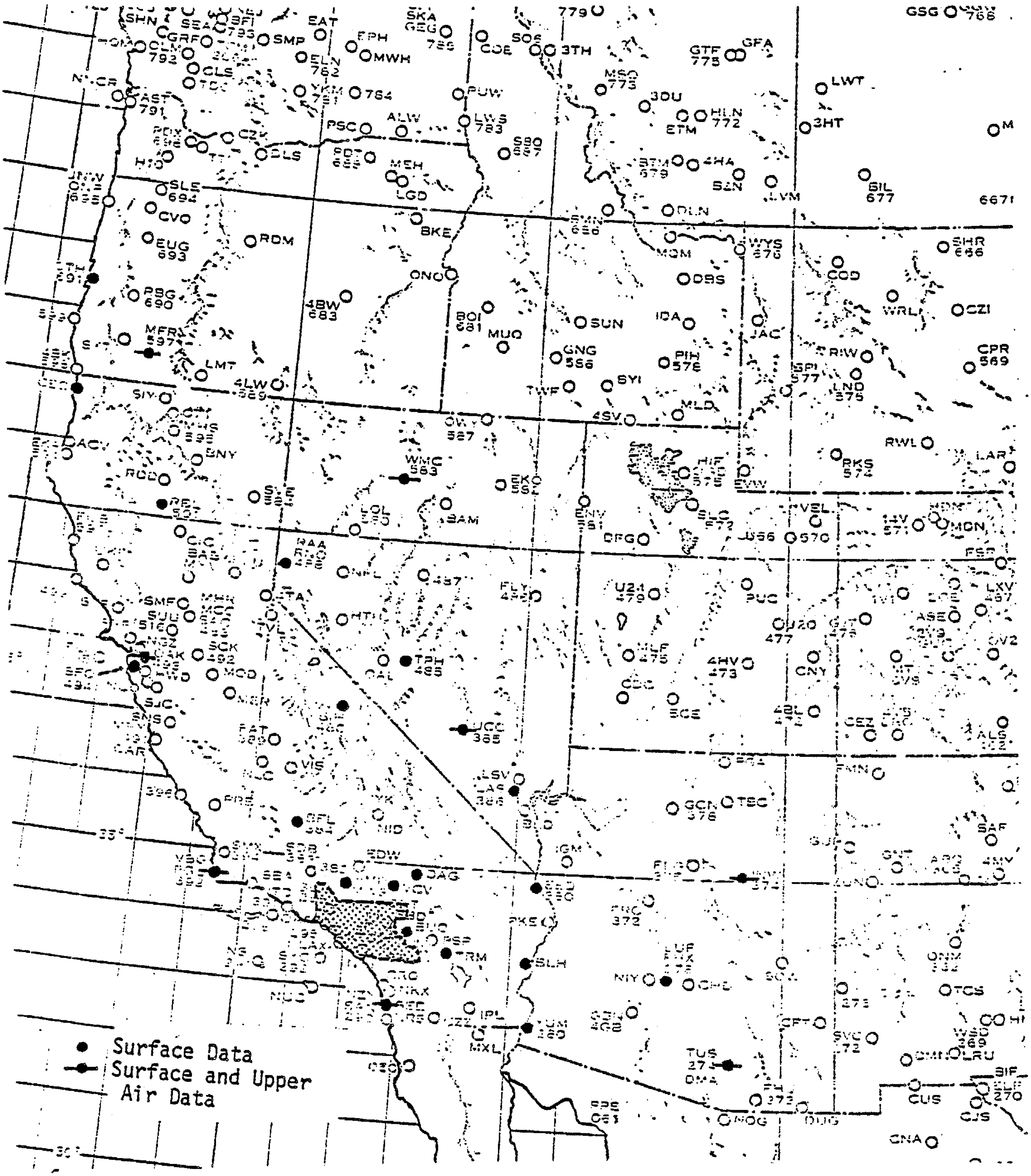


Figure 2.9 Location of Meteorological Sites

were extracted for various valid times 4 A.M., 7 A.M., 1 P.M., 4 P.M., and 10 P.M.

Historical upper air data were assembled for 10 stations located throughout the western U.S. (see Figure 2.9.) Variables were taken from these upper air stations at four different levels 950 mb, 850 mb, 700 mb and 500 mb and included temperature, height, 24 hour height change, wind direction and speed, relative humidity, and temperature anomalies. (Variables were not extracted for all levels or at every station.) Local inversion (RAOB) variables were valid at 14Z and 20Z while other stations data were valid at 12Z and 00Z.

Independent Data Set - 1977

The independent 1977 data set was constructed using variables selected by the various final prediction algorithms. Air quality variables included ozone, SO_2, and sulfates, monitored at the selected key stations. Meteorological variables included a combination of local and distant surface and upper air data. Progostic data in the form of NWS numerical simulation model output (LFM 24-hour 500 mb heights and height differences) were also a part of the data base.

Data assembled for 1977 were used for the independent verification analysis of the newly developed forecast algorithms. The completeness of the independent data set relied heavily upon the availability of the 1977 data.

CHAPTER 3

OXIDANT PREDICTION TECHNIQUES

3.1 GENERAL METHODOLOGICAL OVERVIEW

3.1.1 Selection of Key Predictor Sites

During the research period, there were 38 sites in the South Coast Air Basin which measured oxidants (ozone). To develop site specific equations for each location would have required a substantial statistical effort; but more importantly, individual algorithms may have resulted in regional prediction inconsistencies which could only be attributed to the underlying algorithm. Such results would have necessitated corrections by subjective evaluations of the output data. Therefore it was agreed upon at the outset of the project, that algorithms be developed for five key sites (in the SCAB), each encountering maximal oxidant under a different meteorological pattern. Remaining stations can then be related to these five sites by regression techniques.

The five sites selected (as shown in Figure 3.1) are as follows:

(1) Downtown Los Angeles (DOLA) - a metropolitan area in which pollution is carried inland by the seabreeze. Very specific meteorological conditions are necessary to cause high oxidant levels at this site.

(2) Upland (UPLA) - a foothill receptor site which has experienced some of the highest oxidant levels in the SCAB over the last several years. Predictions for UPLA represent a reasonable potential for basin-maximum concentrations.

(3) La Habra (LAHB) - a northern Orange County site which experiences high oxidant levels when meteorological conditions favor stagnant conditions in the Orange County-Los Angeles County major metropolitan source areas. Like DOLA, on most days, the seabreeze advects the pollution farther inland, thus resulting in more persistent low concentrations.

(4) Riverside (RIVR) - an inland receptor site which is affected by transport through the Santa Ana Canyon. When offshore flow forces the major pollution cloud to a more southerly transport trajectory, Riverside can experience some of the highest concentrations in the SCAB. Generally, it

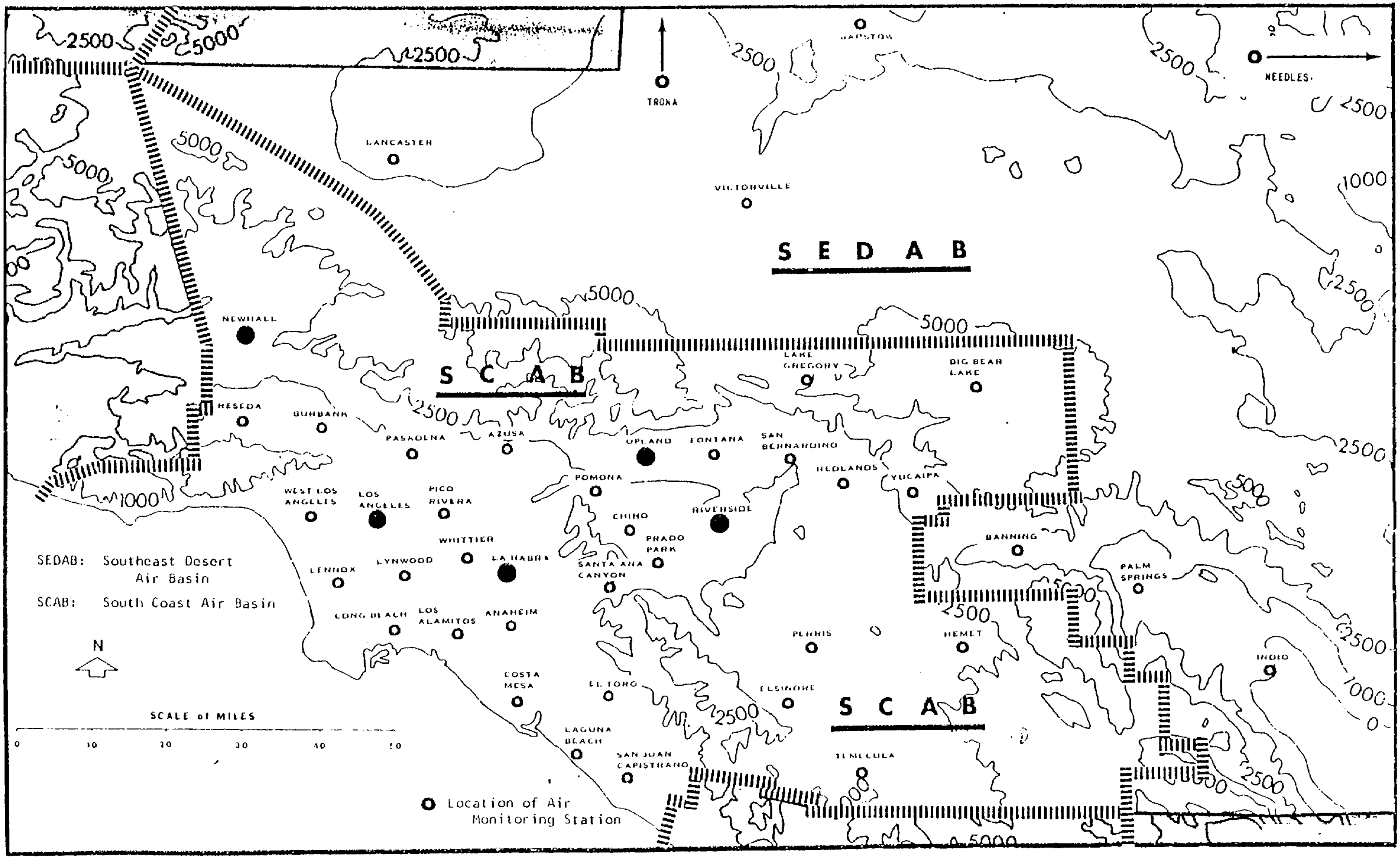


Figure 3.1 Locations of the Five Key Predictor Sites.

can be stated that UPLA represents the maximum potential for "northern route" transport (along the foothills), while RIVR represents the maximum potential for the "southern route".

(5) Newhall (NEWH) - a receptor site in the hills north of the San Fernando Valley which experiences high oxidant levels when reinforced southerly flow (i.e., eddy circulation) exists over Southern California. In situations with a deepening of the marine layer taking place, NEWH can experience the highest oxidant levels in the SCAB.

Since the effects for each of these locations represent distinct meteorological scenarios, it was expected that the relationships (similarities and differences) among these sites would provide enough resolution for predicting values at other SCAB locations. Regression equations were generated for each of the remaining 33 sites, with observed values for the five key sites as independent variables. Separate equations were generated for both summer (May-Oct) and winter (Nov-Apr) periods, however, the resulting equations were in close-enough agreement to warrant using only one set of equations for the entire year. A list of equations (based on the summer data) is presented in Table 3.1.

The resulting equations appear to be physically meaningful, in the sense that each station explained the most variance under a specific group of meteorological effects and transport patterns. Shown in Figures 3.2 to 3.6 are the grouped stations which had the most significant relationship to a particular key site. For each of the five groups, the station with the second most significant influence is shown by a letter immediately adjacent to the monitoring site.

Statistical correlations among stations revealed that the highest correlations occurred geographically between Pasadena and Riverside - the area of the greatest oxidant impact. Figure 3.7 depicts isopleths of correlation coefficient values over the SCAB. The occurrence of high correlations in high oxidant areas, and lower correlations along the coast (in low oxidant areas) resulted in a low standard error of estimate among all stations (approximately 1.5 to 3.0 pphm).

Table 3.1 Prediction Equations for SCAB Sites as Functions of the Five Key Sites

STATION	EQUATION	N	R	R^2	SE
ANAHEIM	ANAH = 0.34 LAHB + 0.30 DOLA + 0.50	802	0.85	0.72	2.12
AZUSA	AZUS = 0.61 UPLA + 0.39 DOLA + 0.09	835	0.95	0.90	2.47
BIG BEAR	BGBE = 0.22 RIVR + 0.09 NEWH - 0.08 LAHB + 2.63	364	0.70	0.49	1.85
BURBANK	BURK = 0.65 DOLA + 0.26 NEWH + 0.16 UPLA + 0.14	835	0.91	0.83	2.72
CHINO	CHIN = 0.47 RIVR + 0.25 LAHB + 0.23 UPLA + 0.49	872	0.94	0.89	2.43
COSTA MESA	COST = 0.26 LAHB + 0.18 DOLA - 0.09 RIVR + 2.46	766	0.62	0.39	2.36
EL TORO	TORO = 0.50 LAHB + 0.12 RIVR + 1.39	805	0.85	0.72	2.35
FONTANA	FONT = 0.65 UPLA + 0.35 RIVR - 0.43	842	0.95	0.90	2.88
HEMET	HEME = 0.28 RIVR + 0.10 NEWH + 2.14	864	0.76	0.58	2.28
LAGUNA BEACH	LGNA = 0.23 LAHB + 0.18 DOLA - 0.08 RIVR + 2.88	219	0.71	0.50	2.04
LAKE ELSINORE	ELSN = 0.40 RIVR + 0.15 UPLA + 2.25	689	0.87	0.75	2.49
LAKE GREGORY	LKGR = 0.41 NEWH + 0.34 UPLA - 0.19 DOLA + 2.25	524	0.77	0.59	3.58
LENNOX	LEXN = 0.36 DOLA - 0.09 UPLA + 2.29	828	0.57	0.33	2.00
LONG BEACH	LONB = 0.20 DOLA + 0.09 LAHB + 1.23	828	0.64	0.41	1.82
LOS ALAMITOS	LSAL = 0.35 LAHB + 0.26 DOLA + 1.58	802	0.81	0.66	2.33
LYNWOOD	LYND = 0.41 DOLA - 0.08 NEWH + 0.10 LAHB + 1.76	803	0.72	0.52	2.31
MT. LEE	MTLE = 0.82 DOLA + 0.19 NEWH + 0.08	252	0.85	0.72	3.19
PASADENA	PASD = 0.42 UPLA + 0.63 DOLA + 0.42	835	0.94	0.88	2.51
PERRIS	PERI = 0.51 RIVR + 0.25 NEWH + 1.56	860	0.90	0.81	2.52
POMONA	POMA = 0.50 UPLA + 0.34 RIVR - 0.46	828	0.96	0.92	2.20
PRADO PARK	PRPK = 0.55 RIVR + 0.27 LAHB + 1.35	841	0.93	0.86	2.27
REDLANDS	REDL = 0.57 RIVR + 0.26 UPLA + 0.44	879	0.94	0.89	2.37
RESEDA	RESD = 0.46 NEWH + 0.34 DOLA + 0.23 RIVR + 0.26	835	0.93	0.86	2.38
RIVERSIDE MAGNOLIA	RIVM = 0.78 RIVR + 0.08	871	0.96	0.91	1.83
SAN BERNARDINO	SNBD = 0.50 RIVR + 0.40 UPLA - 0.37	874	0.94	0.88	2.71
SAN JUAN CAPISTRANO	SJCA = 0.28 LAHB + 0.13 DOLA + 2.51	758	0.63	0.40	2.80
SANTA ANA CANYON	SACN = 0.62 LAHB + 0.36 RIVR + 0.23	571	0.90	0.82	2.64
TEMECULA	TEME = 0.24 RIVR + 0.10 LAHB + 2.94	781	0.77	0.60	2.95
TEMPLE CITY	TEMC = 0.53 UPLA + 0.58 DOLA - 2.12	503	0.95	0.90	2.43
WEST LOS ANGELES	WEST = 0.64 DOLA - 0.08 UPLA + 2.16	828	0.81	0.61	2.09
WHITTIER	WHTR = 0.51 LAHB + 0.38 DOLA + 0.42	884	0.85	0.72	3.00
YUCAIPA	YUCI = 0.58 RIVR + 0.28 NEWH + 0.80	489	0.92	0.84	2.50

N ≡ number of cases

R ≡ correlation coefficient

R^2 ≡ percent of variance explained

SE ≡ standard error of regression

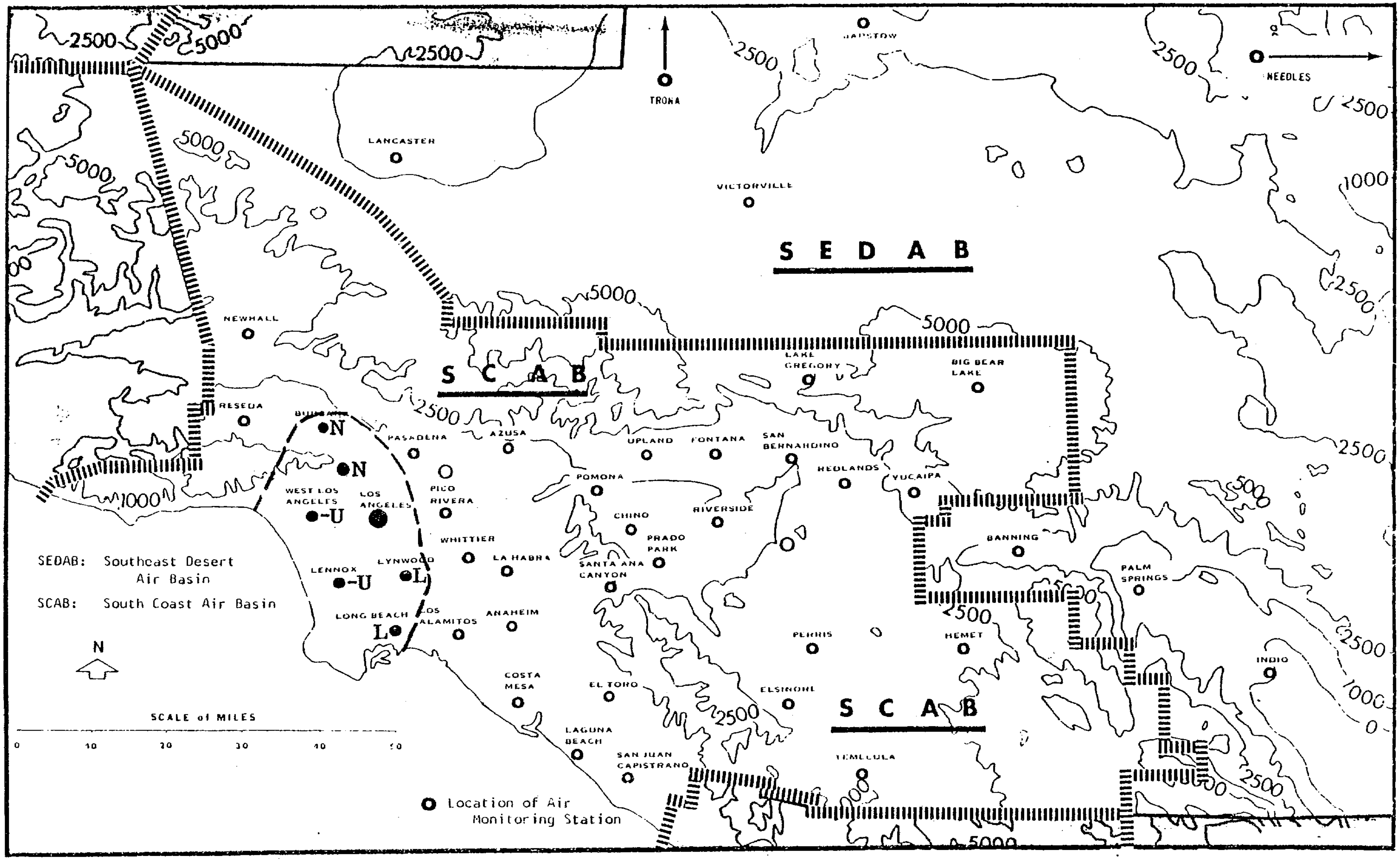


Figure 3.2 Stations with Most Significant Relationship to DOLA. Stations with Secondary Importance are N=Newhall, U=Upland, and L-La Habra.

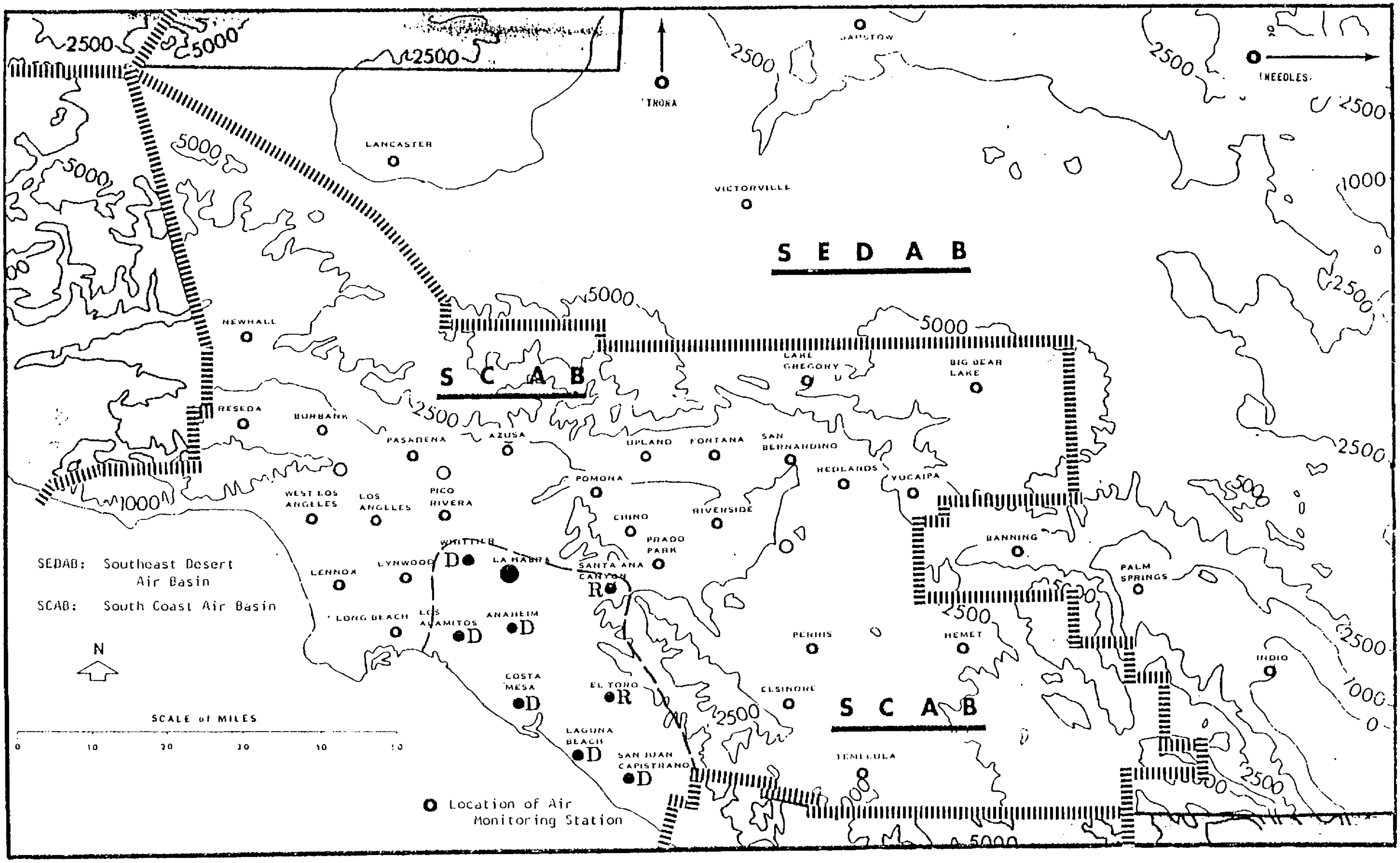


Figure 3.3 Stations with Most Significant Relationship to La Habra.
Stations with Secondary Importance are D=DOLA, R=Riverside.

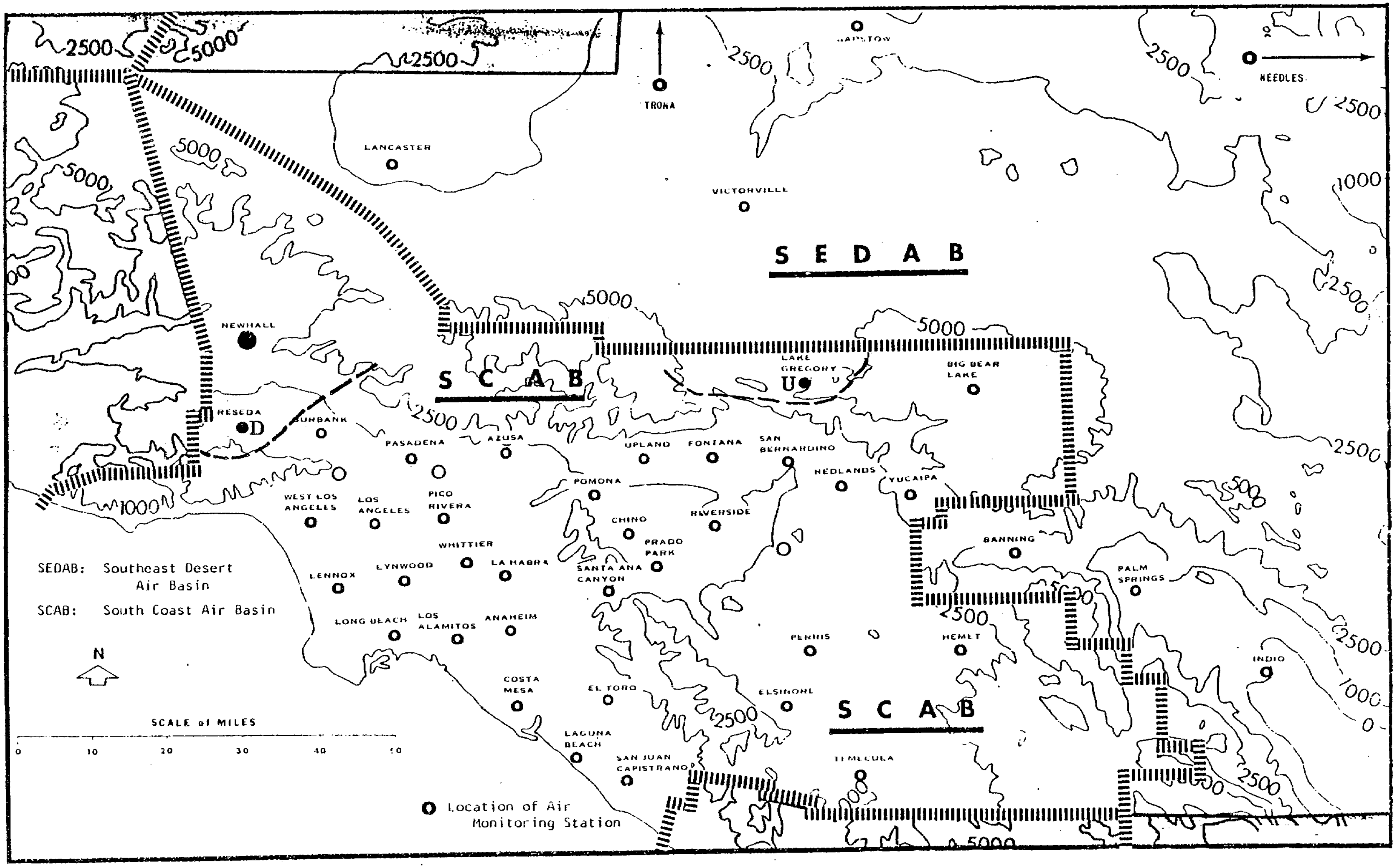


Figure 3.4 Stations with Most Significant Relationship to Newhall. Stations with Secondary Importance are D=DOLA, U=Upland.

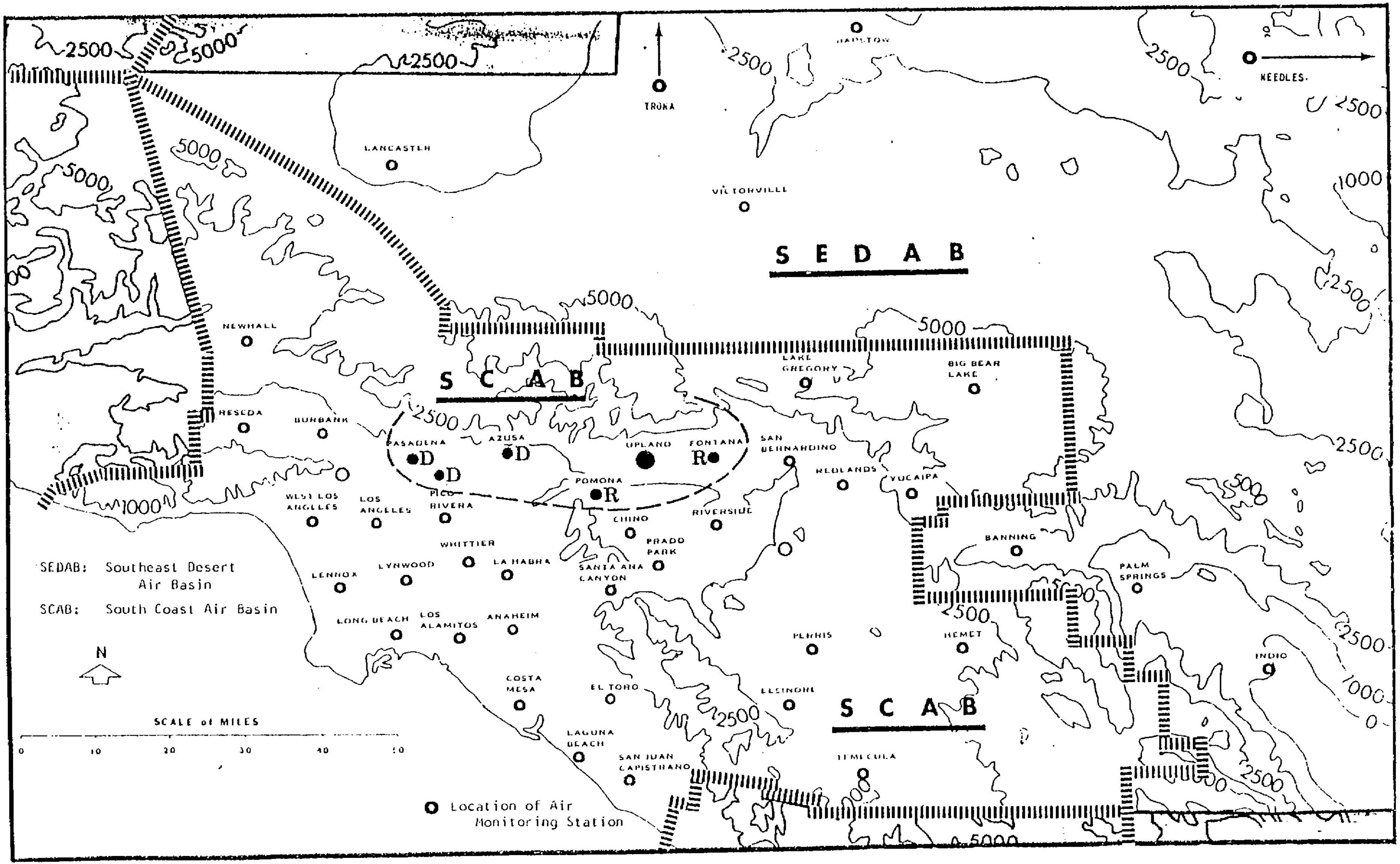


Figure 3.5 Stations with Most Significant Relationship to Upland.
Stations with Secondary Importance are D=DOLA, R=Riverside.

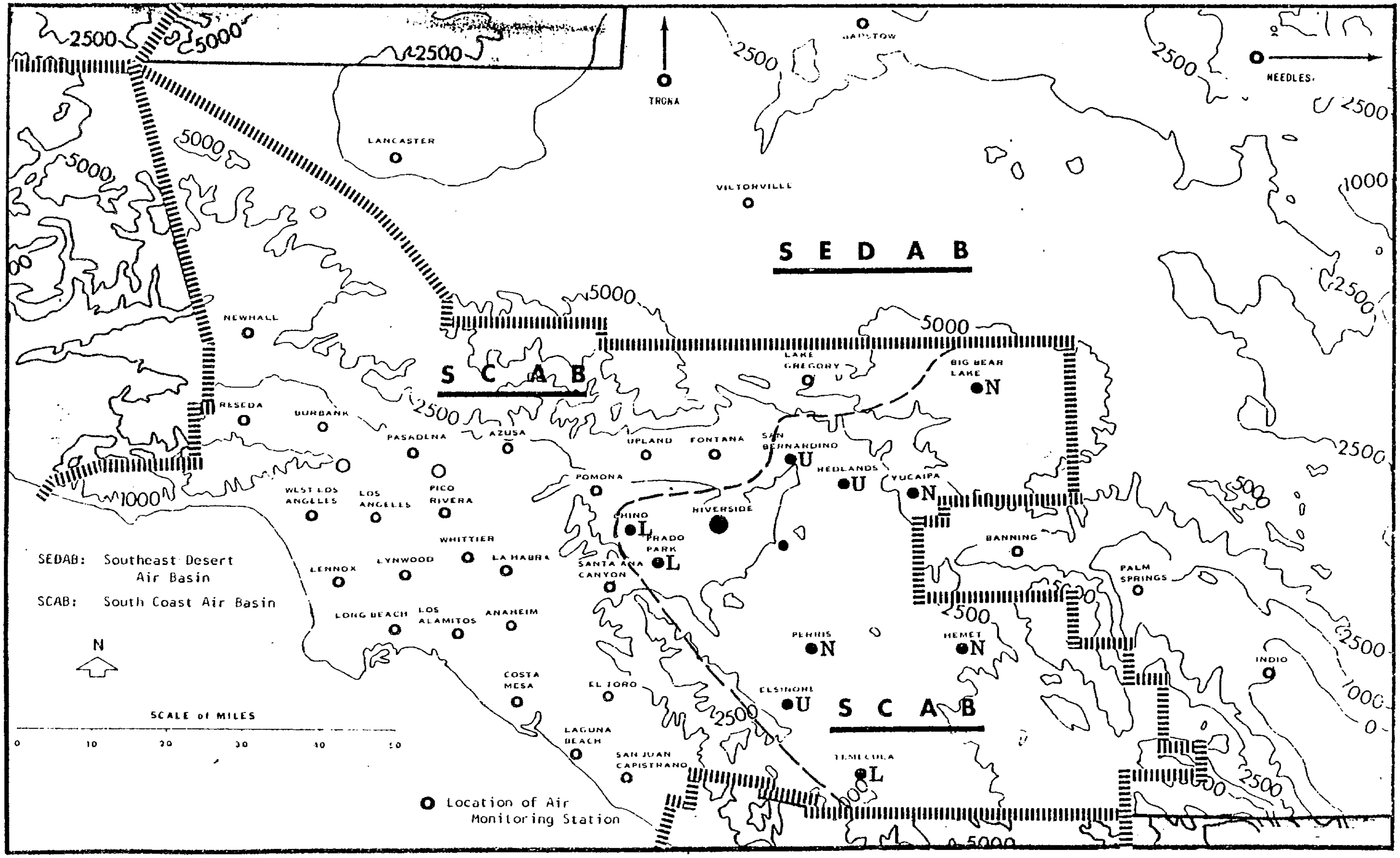


Figure 3.6 Stations with Most Significant Relationship to Riverside. Stations with Secondary Importance are L=La Habra, U=Upland, and N=Newhall.

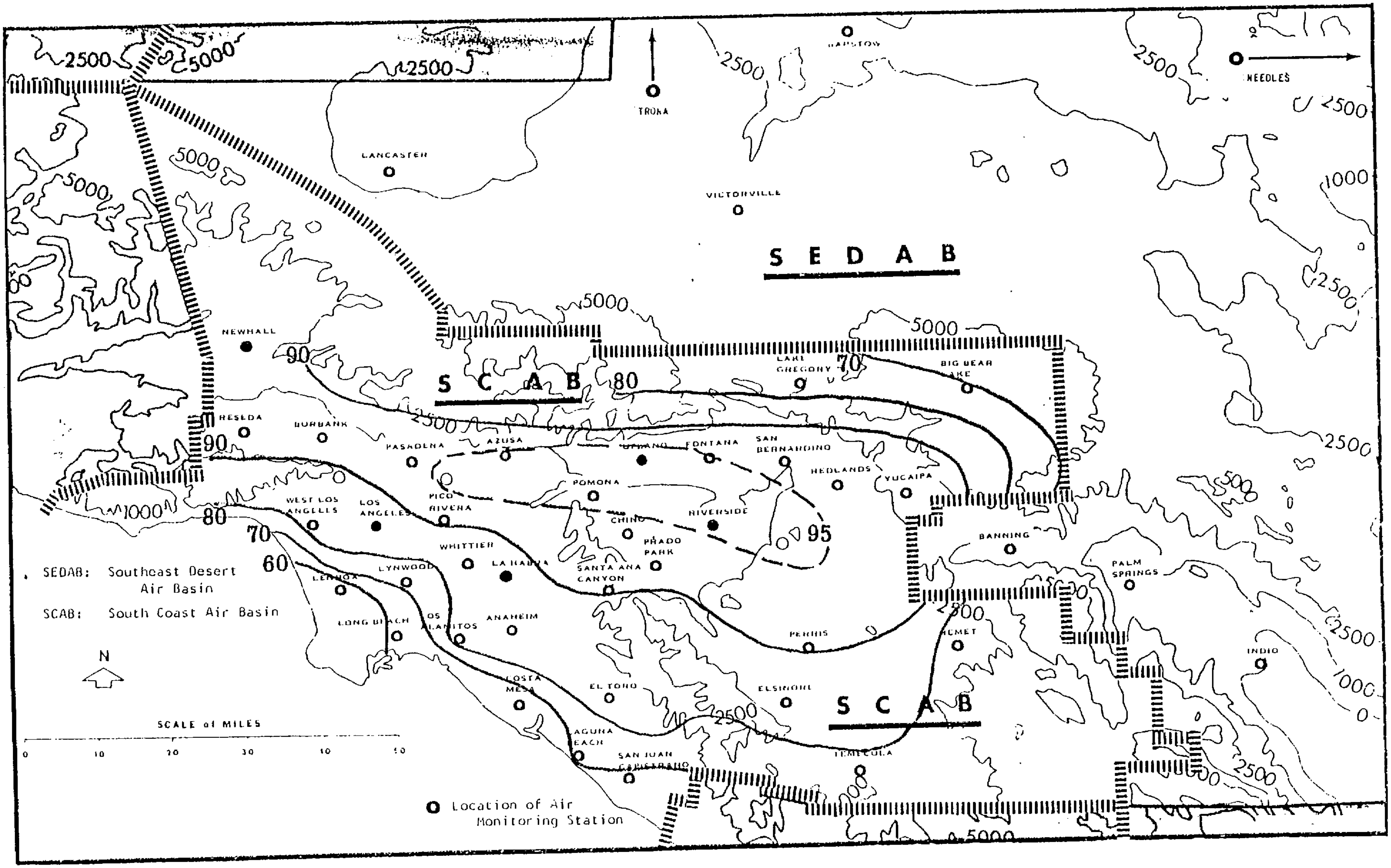


Figure 3.7 Isopleths of Correlation Coefficient Values.

3.1.2 Description of Statistical Techniques Employed

To obtain the greatest predictive capability, many statistical techniques were used, including:

- stepwise multiple linear regression
- automatic pattern recognition ("AID")
- scatterplot analysis
- time series
- nearest neighbor
- combinations of the above techniques
- interactive analysis

The following paragraphs summarize the methodology involved in each of these techniques:

- Stepwise Multiple Regression.

Stepwise multiple linear regression is a statistical process where a dependent variable is fit against a series of predictors having different potential weights,forming a regression equation. If x_1, $x_2,\ldots,x_N$ are N proposed predictive variables, then stepwise multiple regression analysis attempts to find which of these variables, say x_1, x_3, x_4, and x_8, best predict the dependent variable y and to find the optimal prediction equation, say

$$y = H[x_1, x_3, x_4, x_8]$$

where H is a linear function and the variables are the best linear predictors.

The resulting regression equation is a "best fit" of the predictors to the dependent variables. One of the advantages of using stepwise regression is that a definitive set of optimal predictors is formed and that the order of significance is clearly stated. The prediction model that results from the analysis will perform with a respectable amount of accuracy; however, the "best fit" approach frequently underpredicts the high concentrations. Applications of weighted regression (placing emphasis on the high end) can produce a slightly better fit at the high end but will overpredict moderate and low values, thereby increasing the standard error over a non-weighted regression. Problems are also encountered when non-linearly related variables are entered into the set of predictors producing an unrealistic predictor-predictant relationship.

- Automatic Pattern Recognition

An alternative prediction method involves the use of automatic pattern recognition, "AID", to determine pollution decision-trees. The principle behind AID is to categorize a daily pollution level according to meteorological criteria. This is accomplished through the splitting of a selected set of meteorological parameters (each of which are classified in discrete intervals) by maximizing the sum of the squares between the meteorological classes. The AID computer program, developed at the Institute for Social Research, University of Michigan, acts to maximize the sum of the squares:

$$RSS_b = RSS_o - RSS_w$$

where RSS_o: residual sum of the squares for all cases

RSS_w: residual sum of the squares within each meteorological class

RSS_b: residual sum of the squares between meteorotogical classes

by dividing the set of pollutant concentrations into two meteorological regimes.

The logic of AID is briefly explained below:

Look at any one independent variable, say the jth, having values $x_1,\ldots,x_n$ with corresponding y values $y_1,\ldots,y_n$. The RSS_o before any splitting is defined as

$$RSS_o = \sum_{1}^{n} (y_j - \overline{y})^2.$$

The residual sum of squares for the two groups is

$$RSS_1 = \sum_{y_j \varepsilon G_1} (y_j - \bar{y}_1)^2$$

$$RSS_2 = \sum_{y_j \varepsilon G_2} (y_j - \bar{y}_2)^2 .$$

As shown in Appendix E, the decrease in RSS due to the split is

$$\Delta RSS = n_1(\bar{y}_1)^2 + n_2(\bar{y}_2)^2 - n(\bar{y})^2.$$

Consider all partitions of $y_1,\ldots,y_n$ into two groups, and define the optimal split as that partitioning which maximizes ΔRSS.

(1) Pick the variable having the largest ΔRSS, and split the y-values into the two subgroups G_1, G_2, corresponding to this maximum ΔRSS.

(2) Repeat (1) on each of the two groups G_1 and G_2.

Keep repeating the process as long as either :

- a value of ΔRSS exceeds some preset lower bound, or
- The number of elements of a subgroup falls above some present bound.

Results of this method produce a decision-tree, such as the example in Figure 3.8. The AID program was applied to one-day oxidant prediction for Upland, California. The resulting tree defines the most significant meteorological predictor variables, the variable split points, the number, mean, and standard deviations for each tree node, and the identification and rank importance of each terminal node.

- Scatterplot Analysis

Scatterplots were used as prediction aids for two reasons: to present a visual account of the pollutant distribution versus one or two dependent variables and to determine if certain values of an independent variable could be used to isolate regions of similar dependent variable values. Figure 3.9 shows a sample scatterplot for today's oxidant distribution at Upland versus the Vandenberg 500 mb height (12Z).

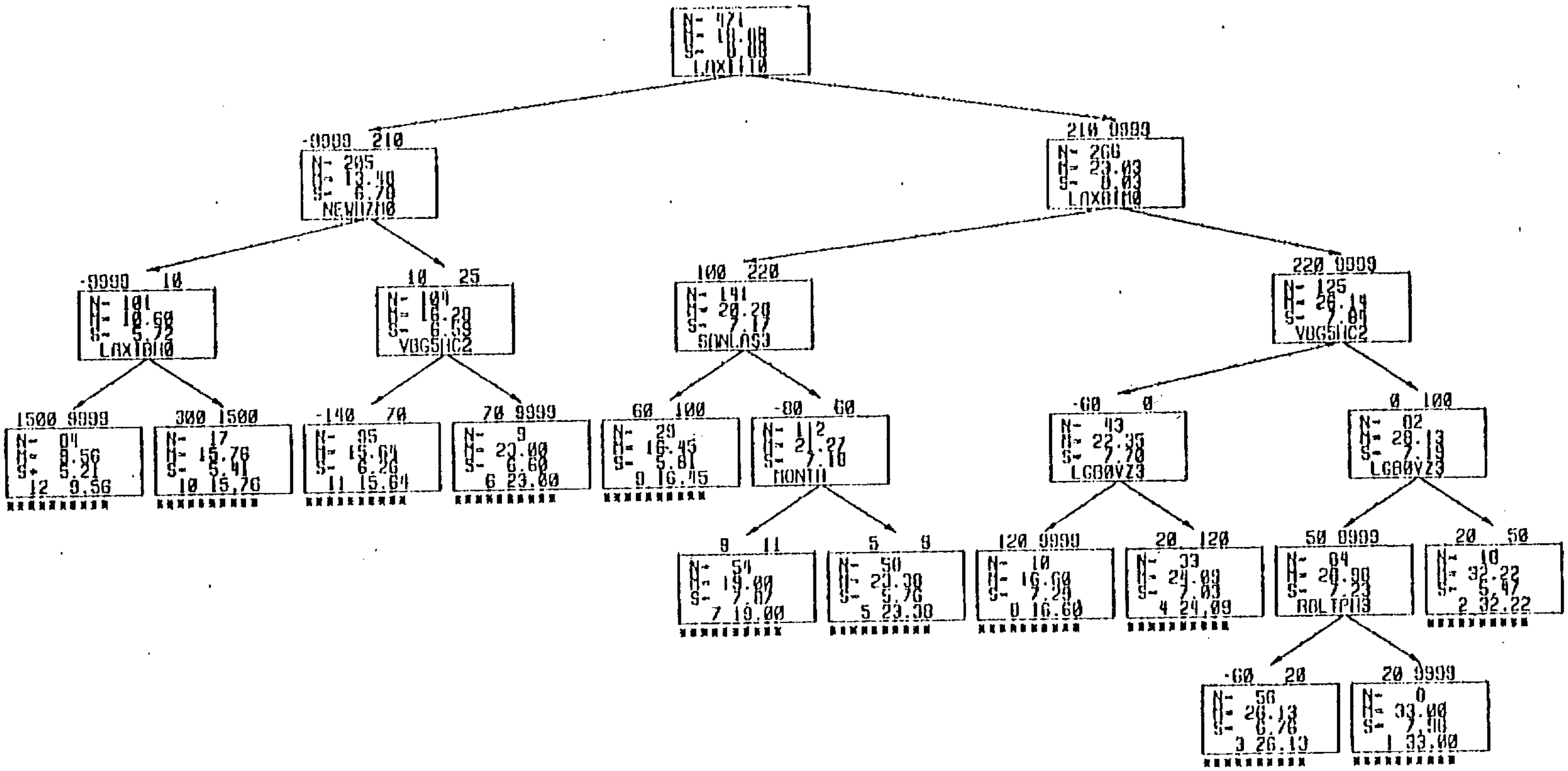


Figure 3.8 Example of Prediction Algorithm Using "AID"

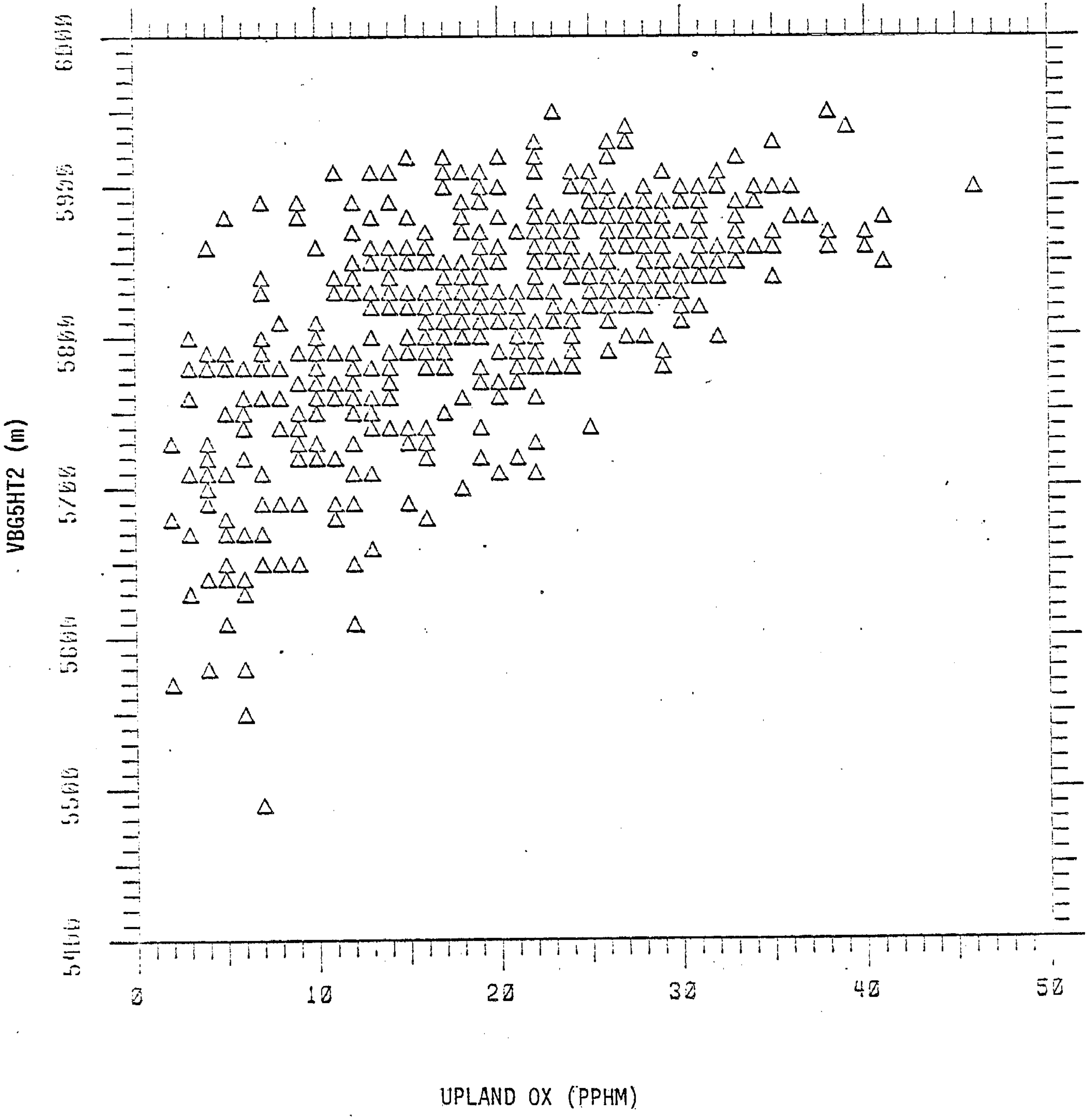


Figure 3.9 Scatterplot of Upland Oxidant Versus Same-Day VBG 500 mb Heights

This technique can be used as follows:

(1) Identify a region, $X_1 < x_1$, such that similar conditions of the dependent variable exist.

(2) For $X_1 > x_1$, plot the remaining values of Y against variable X_2.

(3) Repeat this process until no further reduction for Y is possible.

- Time Series Analysis

The use of a time series model involves the prediction of a future event based solely upon past occurrences of that event. In pollution prediction, time series represents a modified form of persistence - expanding upon persistence by its examination of the recent trend and the general mean of the pollutant forecast. The time series model will overpredict if the immediate observed trend is towards lower concentrations and will reverse itself in the opposite case. The time series can be described as a forecast based upon an observed sequence of a pollutant's concentrations.

A sample representation of a time series equation is shown for a one-day forecast of the DOLA oxidant, where:

$$Y_{t+1} = 1.53y_t - 0.53y_{t-1} - 0.94E_t \qquad (3.1)$$

and $$E_t = y_t - 1.53y_{t-1} + 0.53y_{t-2} + 0.94E_{t-1} \qquad (3.2)$$

Given the past occurrences of y_t, $y_{t-1}, \ldots . y_{t-n}$ a series of residuals E_t, $E_{t-1} \ldots . E_{t-n}$ can be determined. To forecast y_{t+1} (tomorrow's DOLA oxidant) the prediction is dependent upon today's value y_t and the effects of the DOLA oxidant for the past three days - determined by the residuals. In a sense, the residuals damp the forecast, consistently suppressing large scale rapid changes.

- Nearest Neighbor

The nearest neighbor method defines pollution prediction by determining (or "matching") a set of days having meteorological conditions most similar to the conditions occurring at the time the prediction is to be made and then averaging the "matched" pollutant concentrations to form a predicted value.

The basic concept of the nearest neighbor prediction is to represent previous oxidant values (y), and the most current meteorological variables (x_n) in vector form where:

$\overset{>}{x} = (y, x_1, x_2, x_3, \ldots x_i)$ is today's vector and

$\overset{>}{x}^*_m = (y^*, x_1{}^*, x_2{}^*, x_3{}^*, \ldots . X_i{}^*)$ is the vector for days having similar conditions, $m = 1 \ldots n$.

The nearest neighbor technique determines the vector distance $d^{>}(\overset{>}{x}, \overset{>}{x}^*_m) = d\ (\overset{>}{x}, \overset{>}{x}_1{}^*),\ d\ (\overset{>}{x}, \overset{>}{x}_2{}^*) \ldots d\ (\overset{>}{x}, \overset{>}{x}^*_m)$ between today's vector (previous oxidant and meteorological conditions) and the vectors of other days. Once the set of days having the smallest differences (most similar conditions) to today's vector are determined, an oxidant prediction can be made, based upon the oxidant values observed on those days. This is accomplished by calculating the average oxidant value of the set of nearest neighbors. Also determined are the range of oxidant values including the max, min and median values of OX in the nearest neighbor set.

For this analysis the five nearest neighbors were determined for oxidant prediction. The mean, the median and the maximum oxidant values were each examined as potential oxidant predictors.

Another approach for nearest neighbor techniques involves regression. To find the predicted value for a specified case, the k nearest neighbors (the cases whose values of the independent variables are closest) are chosen and then the k values of the dependent variable are either averaged or fit to a linear regression equation. Both the average and the linear nearest neighbor estimator methods weight the closer neighbors more heavily than those more distant. Analyses for several values of k can be done at one time. For each analysis the root mean square (RMS) error of prediction is calculated and is used to compare the two estimators and to aid in selecting a suitable value of k for either estimator. The RMS error is also useful to compare these estimators with the estimates from a linear fit over the whole region.

The three estimators can be defined as follows:

(1) Linear estimator

Let Y denote a dependent variable and $X_1, \ldots, X_\ell$ denote the independent variables. Let $X_{1i}, \ldots, X_{\ell i}, Y_i$ be the values of the variables $X_1, \ldots, X_\ell$, Y for the ith case. Choose $\hat{a}$, $\hat{b}_1, \ldots, \hat{b}_\ell$ to solve the least squares equations that correspond to minimizing the sum of squares

$$\sum_i (Y_i - a - b_1 X_{1i} - \ldots - b_\ell X_{\ell i})^2 \qquad (3.3)$$

Let $X_{1p}, \ldots, X_{\ell p}$ denote the values of the independent variables $X_1, \ldots, X_\ell$ corresponding to the predicted case. The linear estimator of the regression function of Y evaluated at the predicted case is defined and computed as

$$\hat{a} + \hat{b}_1 X_{1p} + \ldots + \hat{b}_\ell X_{\ell p} \qquad (3.4)$$

(2) Weighted average nearest neighbor estimator

Let Y denote a dependent variable and let k be the number of nearest neighbors to be used. For $1 \leq i \leq k$, set $W_i = (k+1)^2 - i^2$ and let Y_i be the value of the dependent variable Y for the case corresponding to the ith smallest distance to the predicted case. The K weighted average nearest neighbor estimator of the regression function of Y evaluated at the predicted case is defined and computed as

$$\frac{\sum_{i=1}^{k} W_i \, Y_i}{\sum_{i=1}^{k} W_i} \, . \qquad (3.5)$$

(3) Weighted linear nearest neighbor estimator

Let Y denote a dependent variable, let $X_1,\ldots,X_\ell$ denote the independent variables and let k be the number of nearest neighbors to be used. For $1 \leq i \leq k$, set $W_i = (k+1)^2 - i^2$ and let $X_{1i},\ldots, X_{\ell i}, Y_i$ be the values of the variables $X_1,\ldots, X_\ell$, Y for the case corresponding to the ith smallest distance to the predicted case. Choose $\hat{a}, \hat{b}_1,\ldots, \hat{b}_\ell$ to solve the least squares equations that correspond to minimizing the weighted sum of squares

$$\sum_{i=1}^{k} W_i(Y_i - a - b_1X_{1i} - \ldots - b_\ell X_{\ell i})^2 \ . \qquad (3.6)$$

Let $X_{1p},\ldots, X_p$ denote the values of the independent variables $X_1,\ldots,X_\ell$, corresponding to the predicted case. The k weighted linear nearest neighbor estimator of the regression function of Y evaluated at the predicted case is defined and computed as

$$\hat{a} + \hat{b}_1X_{1p} + \ldots + \hat{b}_\ell X_{\ell p} \ . \qquad (3.7)$$

- Combination of the Above Techniques

With numerous statistical techniques available, several options were open to determine the most effective prediction algorithm. One of these options was to combine different techniques.

The use of AID and stepwise multiple linear regression was widely used to determine the combined effects of linearly and non-linearly related variables on a pollutant. Using AID, both types of variables are used to separate pollutant concentrations into different categories. With the different categories having discrete characteristics it can be advantageous to group several classes together and then perform regression analyses on those groups to further clarify prediction resolution. The use of regression added to AID also gives the algorithm a continuous prediction capability.

AID was also used to examine the residuals of several regression analyses to increase prediction resolution. Using the combined process, only a small amount of additional variance in the pollutant distribution was accumulated.

- Interactive Analysis

Interactive analysis incorporated the use of personal expertise and empirical data analysis. Several different procedures were attempted to determine predictive algorithms. Of these, point classification systems and unique combinations of key variables (obtained by classical methods) were instrumental in the development of several predictive algorithms.

Point classification systems, similar to Zeldin and Thomas (1975) assigned weights to differing values of selected variables to determine optimal fits of predictor data to forecast pollutants. This can be viewed as a special-case regression method.

3.1.3 Exploratory Meteorological Analyses

Prior to performing statistical analyses between air quality and meteorological parameters, some preliminary analyses among meteorological parameters were accomplished in order to:

(1) determine the extent to which known key meteorological parameters can be statistically predicted, and

(2) determine if upstream parameters are important as potential predictors.

Historically, it has been known that the inversion base height at LAX is an important parameter in determining oxidant concentrations. Therefore this parameter was selected as the predictand for exploratory analysis. Stepwise regression techniques were employed, using pre-selected meteorological predictor variables (six local and five upstream), under four stratified conditions: (1) all cases, (2) May-October days, (3) May-October days with inversion $\Delta T \geq 3°C$ (significant inversion days), and (4) May-October significant inversion days with non-surface inversions. Each of the cases is a subset of the preceding case. For comparative purposes, case #3 was repeated using the logs of the predictor variables and also eliminating persistence as a predictor. Results are summarized in Table 3.2. It can be seen that case #3 explained the most variance (60%), however, persistence (LAXIBH4) was the key predictor variable, as was the case for all regressions except when persistence was eliminated. Pressure gradients and pressure gradient changes have a secondary importance in predicting inversion height.

Analyses of the prediction errors indicate that regression techniques fail to adequately predict significant changes in the inversion base height. This accounts for the rather large standard errors (i.e., 810 feet for the best case), and also implies the strong persistency effects. For the purposes of predicting high oxidant values (when meteorological changes are critical) 30 hours in advance, it becomes apparent that best-fit techniques may not provide sufficient resolution to achieve desired results.

Using the same predictor variables, a correlation matrix was generated (see Table 3.3). As expected, local predictor variables are correlated to each other, and most upstream predictors are correlated to each other. Looking at the correlations for predicting tomorrow's inversion base height

Table 3.2 Comparative Statistics for Regressions Predicting LAX Inversion Base Height

	Condition	Correlation Coefficient	Percent Variance Explained	Standard Error (feet)	Number of Cases	Key Predictor Variables	Percent Variance Explained
(1)	All cases	0.63	40%	1389	882	LAXIBH4 PLTOPC7 SUMØPG7 VBG5HT2 others	24% 9% 3% 2% 2%
(2)	May - October	0.67	46%	1055	492	LAXIBH4 PLTOPC7 VBG5HT2 others	39% 5% 1% 1%
(3)	May - October and ΔT ≥ 3°C	0.77	60%	810	373	LAXIBH4 PLTOPC7 SUMØPG7 others	53% 3% 1% 3%
(3a)	Same as above, using logs of met predictors	0.70	49%	914	373	LAXIBH4 VBG5HT2 PLTOPC7 others	40% 5% 2% 2%
(3b)	Same as 3, eliminating persistence as predicator variable	0.69	47%	923	373	SUMØPG7 VBG5HT2 NWCOPC7 others	36% 6% 2% 2%
(4)	May - October and ΔT ≥ 3°C and no surface inversions	0.76	58%	796	333	LAXIBH4 PLTOPC7 SUMØPG7 others	50% 5% 2% 1%

Table 3.3 Correlation Matrix of Predictor Variables

	PREDICTAND	LOCAL PREDICTORS						UPSTREAM PREDICTORS				
	LAXIBH4 (tomorrow)	LAXIBH4	LAX8AN4	SUMØPG7	VBG5HT2	SAN7SD2	LAXIDT4	OAK8TC2	NWCOPC7	PLTOPC7	SFORNO7	OMW7TG2
LAXIBH4 (tomorrow)		0.73*	-0.37*	0.61*	-0.47*	-0.13*	-0.16*	-0.27*	0.08	0.17*	0.28*	0.06
LAXIBH4			-0.54*	0.67*	-0.57*	-0.15*	0.59*	-0.24*	-0.06	-0.01	0.24*	0.08
LAX8AN4				-0.42*	0.42*	0.33*	-0.13*	0.18*	0.17*	0.16*	-0.17*	0.10
SUMØPG7					-0.41*	-0.27*	0.44*	-0.40*	-0.11	0.17*	0.44*	-0.02
VBG5HT2						0.21	0.26*	0.22*	0.05	-0.00	-0.33*	0.07
SAN7SD2								0.21*	0.07	-0.12	-0.24*	0.21*
LAXIDT4								0.10*	0.08	0.13	-0.12	0.05
OAK8TC2									0.18*	-0.24*	-0.39*	0.17*
NWCOPC7										0.58*	-0.03	-0.03
PLTOPC7											0.22*	-0.30*
SFORNO7												-0.16*
OMW7TG2												

*Correlations significant at the 99% level

(LAXIBH, tomorrow), it is interesting to note that three of the five upstream predictors did show significant correlations, with SFORNO7 (San Francisco-Reno pressure gradient) and OAK8TC2 (24-hour 850 mb temperature change at Oakland) being the most significant. While not as high as some local predictor variables, these correlations did indicate that upstream meteorological predictor variables may be useful in longer-range oxidant prediction.

3.1.4 Oxidant Correlations to Data Base Variables

As a preliminary analysis to determine sets of potential predictors, linear correlation coefficients between every variable in the data base and the oxidant values at each key station were calculated. Coefficients for variables compared to same day oxidant and stratified by season are given in Appendix B.

Although highly correlated variables may not have been selected for the final set of predictors, they represented a logical starting point for predictor selection.

3.2 SAME-DAY (6-HOUR) PREDICTIONS

Same-day predictions, which in an operational sense can be considered as improved updates of previously issued forecasts, are based on morning observations of meteorological and air quality data. Since the most statistically significant relationships would normally be expected on same-day conditions, our initial efforts were focused on maximizing these results. Final same-day prediction algorithms for the five key sites are presented below. (Details of the chronological development of these algorithms are given in the subsequent section.)

3.2.1 Final Prediction Algorithms

(1) UPLAND

$$\boxed{\text{OX} = \text{LAX8TM4} + \text{LAX8DIF} + (\text{UPLAZMY} - \text{LAX9TM4})} \qquad (3.8)$$

where: LAX8TM4 = LAX 850mb Temp (^{o}C) 14Z

LAX8DIF = Change in LAX 850mb Temp from 14Z yesterday to today (^{o}C)

UPLAZMY = Upland max oxidant yesterday (PPHM)

LAX9TM4 = LAX 950 mb temp (^{o}C) 14Z

Note: (A) $-7.5 \leq (\text{UPLAZMY} - \text{LAX9TM4}) \leq 7.5$

(B) If LAX 950mb Temp from 14Z yesterday to 14Z today (LAX9DIF) is:

(a) > 7.0, then $\text{OX}_1 = 1.1\ (\text{OX})$

(b) <-7.0, then $\text{OX}_1 = 0.6\ (\text{OX})$

(2) RIVERSIDE

$$\boxed{\text{OX} = \text{LAX8TM4} + (\text{LAX8DIF} + \text{LAX9DIF})} \qquad (3.9)$$

where: $\left.\begin{array}{l}\text{LAX8TM4}\\ \text{LAX8DIF}\\ \text{LAX9DIF}\end{array}\right\}$ as defined for Upland

Note: (A) $-10.0 \leq (\text{LAX8DIF} + \text{LAX9DIF}) \leq 10.0$

(B) For October - April:

$$\text{OX}_1 = \frac{\text{OX} + \text{RIVRZMY}}{2}$$

where RIVRZMY = Riverside max oxidant
yesterday (pphm)

(3) NEWHALL

$$\boxed{\begin{array}{l}\text{OX} = 0.35\ (\text{LAX8TM4}) + 0.43\ (\text{NEWHZMY}) \\ \quad + 0.46\ (\text{LAX8DIF}) - 0.25\ (\text{LAX9DIF}) + 1.48\end{array}} \tag{3.10}$$

where: NEWHZMY = Newhall max oxidant yesterday (pphm)

$$\left.\begin{array}{l}\text{LAX8TMY}\\ \text{LAX8DIF}\\ \text{LAX9DIF}\end{array}\right\}\text{ as defined for Upland}$$

(4) DOWNTOWN LOS ANGELES - (decision tree - see Figure 3.10)

(5) LA HABRA - (point - score system - see Figure 3.11)

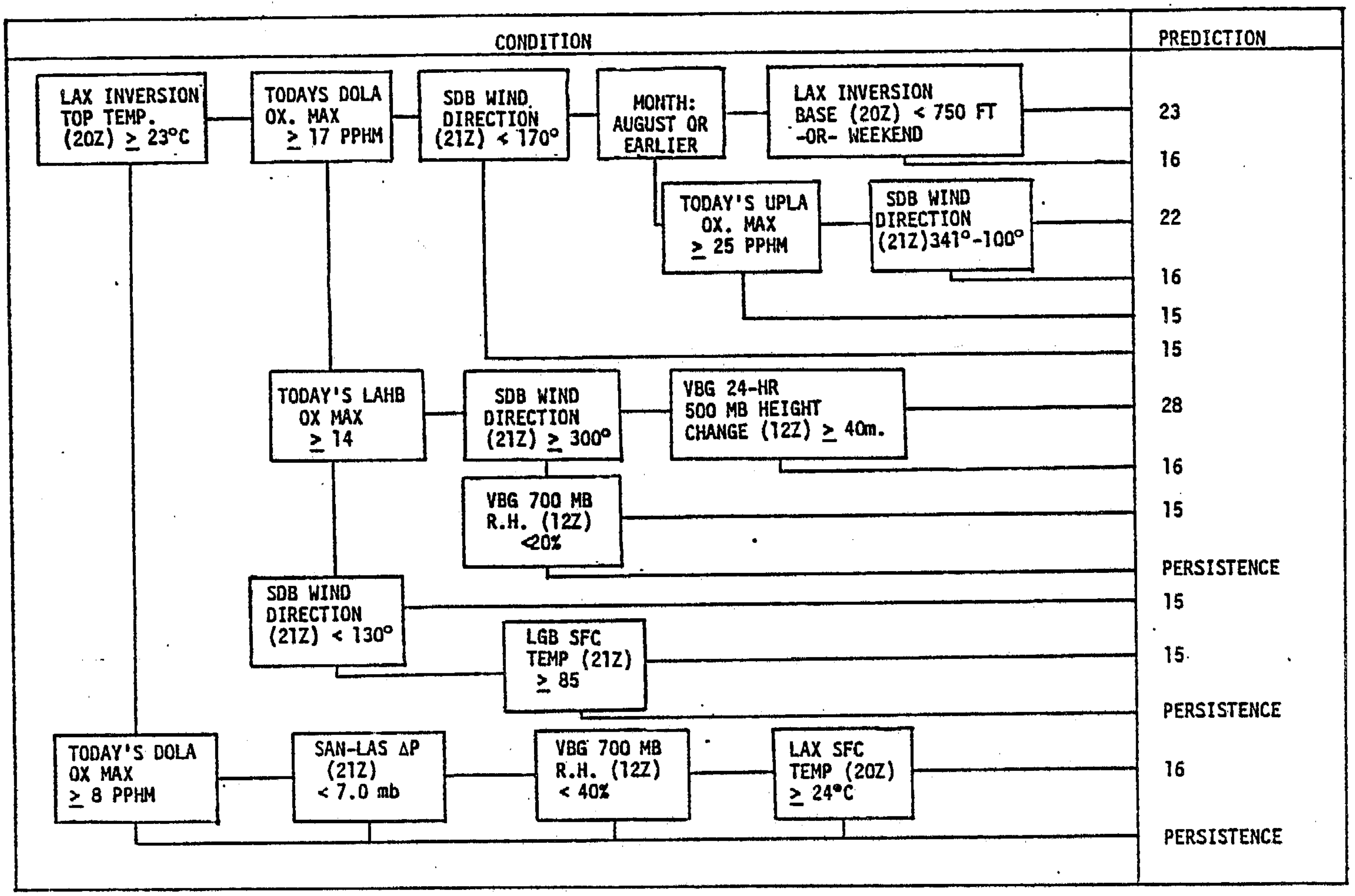

CONDITION
PREDICTION
LAX INVERSION TOP TEMP. (20Z) ≥ 23°C
TODAYS DOLA OX. MAX ≥ 17 PPHM
SDB WIND DIRECTION (21Z) < 170°
MONTH: AUGUST OR EARLIER
LAX INVERSION BASE (20Z) < 750 FT -OR- WEEKEND
23
16
TODAY'S UPLA OX. MAX ≥ 25 PPHM
SDB WIND DIRECTION (21Z)341°-100°
22
16
15
15
TODAY'S LAHB OX MAX ≥ 14
SDB WIND DIRECTION (21Z) ≥ 300°
VBG 24-HR 500 MB HEIGHT CHANGE (12Z) ≥ 40m.
28
16
VBG 700 MB R.H. (12Z) <20%
15
PERSISTENCE
SDB WIND DIRECTION (21Z) < 130°
15
LGB SFC TEMP (21Z) ≥ 85
15
PERSISTENCE
TODAY'S DOLA OX MAX ≥ 8 PPHM
SAN-LAS ΔP (21Z) < 7.0 mb
VBG 700 MB R.H. (12Z) < 40%
LAX SFC TEMP (20Z) ≥ 24°C
16
PERSISTENCE

Figure 3.10 Decision-tree Prediction Algorithm for DOLA (one day in advance). "Yes" condition proceeds horizontally to the right; "No" condition proceeds downward. Predicted values are in PPHM.

Value	LAX8TM4	LAX8DIF	LAX9DIF	LGBØVZ7	SANLAS7	LAXIBC3
-10						N/A
- 9					≤ -10.0	N/A
- 8						(see note 2)
- 7						N/A
- 6						(see note 2)
- 5		≤ -8.0	≤ -9.0		>8.0	N/A
- 4		-7.9 to -6.0	≤ -8.0		8.0 to 6.1 -9.9 to -9.0	N/A
- 3		-5.9 to -4.0	≤ -7.0		6.0 to 4.1	N/A
- 2		-3.9 to -2.0	≤ -6.0	≥15	4.0 to 3.1	N/A
- 1		-1.9 to -1.0	≤ -5.0	14 to 10	3.0 to 2.1 -8.9 to -8.0	N/A
0	≤ 10.0	-0.9 to 0.0	-5.0 to 5.0	9	2.0 to 1.1	(see note 2)
1	10.1 to 12.0	0.1 to 1.0	N/A	8	1.0 to 0.1 -7.9 to -7.0	N/A
2	12.1 to 13.0	1.1 to 2.0	> 5.0	7	0.0 to -0.9	N/A
3	13.1 to 14.0	2.1 to 3.0	≥ 6.0	6	-1.0 to -1.9 -6.9 to -6.0	N/A
4	14.1 to 15.0	3.1 to 4.0	≥ 7.0	5	-2.0 to -2.9 -5.9 to -5.0	N/A
5	15.1 to 16.0	4.1 to 5.0	≥ 8.0	4	-3.0 to -3.9	N/A
6	16.1 to 18.0	5.1 to 6.0	≥ 9.0	3	-4.0 to -4.9	N/A
7	18.1 to 20.0	>6.0	≥10.0	2		N/A
8	20.1 to 22.0			1		N/A
9	22.1 to 24.0			<1		N/A
10	>24.0					N/A

where: LAX8TM4
LAX8DIF as previously noted
LAX9DIF
LGBØVZ7 = LGB Visibility at 0700 PST (miles)
SANLAS7 = (SAN-LAS) surface pressure gradient at 0700 PST (mb)
LAXIBC3 = LAX morning inversion base height change from yesterday to today (feet)

Note: (1) LAX8DIF = 0 for July - October
(2) If LGBØVZ7 ≤ 5 and LAXIBC3 ≥ 0, then point score = -6
If LGBØVZ7 ≤ 5 and LAXIBC3 ≥ 1000, then point score = -8 (For all other conditions, point contribution of LAXIBC3 = 0)

Figure 3.11 Point - Score Predictive Algorithm for La Habra

3.2.2 Chronological Development of Algorithms

The five key sites were studied as separate cases; however, the information gained from certain analytical methods applied to one site was used to prevent unnecessary repetitions on other sites. For example, exploratory research using "nearest neighbor" techniques was applied to DOLA. When it became apparent that the range of possible values under similar meteorological conditions was too great to provide sufficient prediction accuracy, these procedures were not applied to other sites, although information regarding potential predictors was retained for utilization in other methods. This feedback approach maximized efforts leading to the final algorithms.

DOWNTOWN LOS ANGELES - (DOLA)

Initial same-day prediction methods were focused on DOLA, due to the critical nature of meteorological conditions necessary to produce episode conditions. Since most days have values < 20 pphm, climatology has been shown to be an effective predictor (see Phase I report). The few episodes that occur each year (approximately 10/year) are generally non-persistent; hence the importance of an improved prediction method can be more substantially evaluated by its ability to successfully predict episode days. Results of the initial regressions yielded the largest correlations when stratified by weekday/weekend with the weekend equation providing the best fit (r = 0.75, N = 152). The mean absolute error of 2.7 pphm was better than any procedure evaluated in Phase I; however, as typified by regression analyses, the equation was unable to predict high values ≥ 20 pphm. Under the most adverse meteorological conditions of the dependent data set, the highest predicted value was 19 pphm. For the weekday data, results were slightly worse (r = 0.70, N = 361) with a mean absolute error of 3.0 pphm. Also, very poor accuracy in predicting episode days was observed. Key predictor variables are given in Table 3.4.

Next we applied the AID program to a list of 21 potential predictors. Even though 61% of the variance was explained, the decision tree did not yield any categories which predicted ≥ 20 pphm. The highest category,

Table 3.4 Key Predictor Variables for DOLA Same-Day Regression

Weekday		Weekend	
Variable	Coefficient	Variable	Coefficient
LAXITT4	0.39	LAXITT4	0.18
LAX9TM4	0.25	LAX9TM4	0.25
LGBØVZ7	-0.18	LGBØV27	-0.29
VBG5HC2	-0.13	SANLAS7	-0.32
Constant	-0.3	Constant	+6.7

Where: LAXITT4 = LAX INVERSION TOP TEMP (°C) -14Z

LAX9TM4 = LAX 950 mb TEMP (°C) - 14Z

LGBØVZ7 = LGB VISIBILITY (MILES) - 07 PST

SANLAS7 = (SAN - LAS) ΔP (mb) - 07 PST

VBG5HC2 = VGB 500 mb HEIGHT CHANGE (10 m) - 12Z which predicted 18 pphm, involved the following criteria:

(1) SDB temperature at 07PST (SDBØTM7) > 55°F

(2) LAX 850 mb temp. at 14Z (LAX8TM4) > 20°C

(3) LAX inversion base height at 14Z (LAXIBH4) < 500 ft.

As will be shown in the subsequent section, results of the day-in-advance AID algorithm for DOLA verified better than any of the same-day methods. This phenomenon is not completely understood, although it is apparent that the meteorological conditions developing on the previous afternoon are most important in relating to high oxidant conditions at DOLA. Of the five key sites, DOLA is the only one which produced better day-in-advance prediction accuracy than same-day accuracy. A description of this algorithm is given in Section 3.3.

UPLAND - (UPLA)

As a starting point, linear regression methods were used. The most significant resulting equation is as follows:

$$\begin{aligned}\text{UPLA} = {} & 0.66\,(\text{LAX8MØ4}) + 0.18\,(\text{DOLAN9Ø}) + 0.32\,(\text{UPLAZMY}) \\ & - 0.45\,(\text{LAX9TM4Y}) - 0.27\,(\text{SUMØPG7}) + 0.29\,(\text{VBG5HT2}) - 148.9 \end{aligned} \tag{3.11}$$

$r = 0.83 \quad N = 451$

Where:

LAX8MØ4 = LAX (850 mb - sfc)$_{TEMP}$ at 14Z

DOLAN9Ø = DOLA 6-9 a.m. max hourly NO_2

UPLAZMY = Yesterday's max oxidant at UPLA

LAX9TM4Y = Yesterday's LAX 950 mb temp at 14Z

SUMØPG7 = Pressure gradients; |(SAN-LAS) + (LGB - DAG) + (SBD-VCV) + 6| at 0700 PST

VBG5HT2 = Vandenberg 500 mb height at 12Z (10's m)

For verification statistics, this equation outperformed the existing ARB same-day equations (which are stratified according to weekday/weekend). For example, for the 1974 - 1976 period, the existing equations hit 54% of all possible episodes, had a mean absolute error of 7.4 pphm, and a total verification score of 46. (See Section 2.3 for a review of verification methods.) The new equation correctly predicted 71% of all episodes, had a mean absolute error of 4.8 pphm and a total verification score of 105. This, however, was still not as good as the existing AQMD objective system, which scored 114. Also, the new equation did not predict any stage 2 condition (e.g., the maximum predicted value was 33 pphm).

We next proceeded to diagnose the terms of the existing and new equations on a case-by-case basis to determine the conditions in which the equations do not perform well. One of more interesting features we observed was that the DOLA 6-9 a.m. NO_2 maximum hourly value, while important overall, did not seem to correlate under the more severe oxidant cases. This could be attributed to slight wind change conditions which could move the NO_2 "cloud" away from the DOLA station before high values were detected. We sought, therefore, to determine an NO_2 "potential" based on meteorological parameters. Regression analyses indicated that virtually all of the explainable variance (41% of 44%) could be attributed to the LAX 950 mb temperature. Thus, by using that parameter as a surrogate for NO_2 concentrations, we could "stablize" the NO_2 input parameter.

The inclusion of the previous day's LAX 950 mb temperature as a predictor (in the new equation) suggested that the change in key predictor variables could be more important than the actual parameters. By examining the two key change predictors (24-hour changes in the LAX 950 mb temperature [LAX9DIF] and the 850 mb temperature [LAX8DIF]), and by eliminating some of the less significant predictors (i.e. SUMØPG7 and VBG5HT2), we were able to construct a simplified algorithm which is the final algorithm given in Section 3.2.1. On the dependent data set, this method achieved an overall score of 127 -- better than any other method. (See Table 3.11*). The

*Verification tables (3.11 to 3.15) are included in Section 3.5, "Verification," starting on page 111. It was not possible to complete the chronological development of the algorithms without reference, at times, to the verification scores; hence, the reference to such tables may appear out of sequence in the text.

primary improvement in this method over the new regression equation is in its ability to predict stage 2 episodes. The new algorithm successfully predicted 25% of the stage-2 episodes, with only a .005 false alarm rate, whereas equation (3.11) would not predict any stage-2 events. On the independent data verification (Table 3.11), the new method scored a total of 188, which was twice as good as any other method. Of key importance, the method correctly predicted 85% of the episodes, and predicted one out of the two (50%) stage two events with no false alarms. Clearly, it can be seen that no other method even approached this accuracy.

Further improvements of the method were attempted, using best fit techniques of the grouped terms of the algorithm. Both linear regression and weighted linear regression were used, but the results were not as successful as the original algorithm. Hence, this method was selected as the final same-day prediction algorithm for Upland.

LA HABRA - (LAHB)

As in the case for DOLA, La Habra is affected by specific meteorological conditions not necessarily representative of the basin-max conditions. Only when a "southern" route transport trajectory occurs is LAHB susceptible to high ozone concentrations. It is not surprising that straight forward regression techniques were not successful in achieving any reasonable prediction equation. We therefore attempted sets of regression equation, sorted on key variables. The two most productive sets are summarized in Table 3.5. While the best fits occurred with the weekend/weekday stratification, the least standard error occurred with the condition LAXTPH ≥ 0, primarily due to the low observed oxidant concentration induced by onshore flow, represented by the positive gradient. In all cases, stage 2 episodes were not predicted, with a tendency to significantly underpredict high values.

Based on the successful methods developed for predicting oxidant concentrations at UPLA, and based on the regression results which selected key predictor variables, we sought to develop a method which could not only perform well on the low/moderate concentrations, but also for the high oxidant days as well.

Table 3.5 Key Predictor Variables for LAHB Same-Day Regression

Set #1

Weekday		Weekend	
Variable	Coefficient	Variable	Coefficient
UPLAZMY	0.28	LAX9TM4	0.37
SANLAS7	-0.25	LGBØVZ7	-0.33
LAXIBC3	-0.0007	LAXIBC3	-0.0015
LAX9TM4	0.21	UPLAZMY	+0.16
LGBØVZ7	-0.14	LAXTPH7	-0.34
Constant	+2.2	Constant	+3.5
(N = 361, r = 0.65, Se = 4.7)		(N = 148, r = 0.70, Se = 5.3)	

Set #2

LAXTPH7 < 0		LAXTPH7 > 0	
Variable	Coefficient	Variable	Coefficient
LAX8TM4	0.61	LAX8TM4	0.30
LAXIBC3	-0.0012	SANLAS7PC	-0.31
LGBØVZ7	-0.26	LGBØVZ7	-0.21
SANLAS7	-0.50	LAX8DIF	0.17
Constant	+1.4	Constant	+4.3
(N = 241, r = 0.53, Se = 6.2)		(N = 253, r = 0.59, Se = 3.6)	

Where: UPLAZMY = Yesterday's max oxidant at UPLA

SANLAS7 = SAN-LAS pressure gradient at 0700 PST

LAXIBC3 = Change in the LAX inversion base height: 14Z today minus 14Z yesterday

LAX9TM4 = LAX 14Z 950mb temperature

LGBØVZ7 = LGB visibility at 0700 PST

LAXTPH7 = LAX-TPH pressure gradient at 0700 PST

LAX8TM4 = LAX 14Z 850mb temperature

SANLAS7PC = Change in the SAN-LAX pressure gradient; 0700 PST today minus 0700 PST yesterday

LAX8DIF = Change in the LAX 850mb temperature: 14Z today minus 14Z yesterday

An examination of individual meteorological parameters indicated that linear relationships existed over some segment of the variable, but not over the entire distribution. For example, we found that for the LAX 850 mb temperature, values less than 10°C had little influence on oxidant values at LAHB. Similarly, values over 24°C did not significantly affect the high oxidant days. However, between 10°C and 24°C, there was a maximum linear correlation between LAHB oxidant and the 850 mb temperature.

Other parameters which were found to have a segmented linear relationship included: (1) the 24-hour 850 mb temperature change (LAX8DIF), (2) the 950 mb temperature change (LAX9DIF), (3) LGB surface visibility (LGBØVZ7), and the SAN-LAS pressure gradient (SANLAS7). Using a point-score method, the significant portions of these key predictors were combined. The results were presented in section 3.2.1. Verification using this method did substantially better than persistence in both the dependent data set and the 1977 independent data test (See Table 3.13). Also, this method was considerably better than the existing ARB subjective prediction for 1977, which was only slightly better than persistence. Of particular interest in the 1977 verification is that 65% of all predictions were ± 2 pphm, and a respectable 43% of all significant change days were within ± 2 pphm.

From these results, it appeared that a nearest neighbor approach would yield an even better algorithm. Using the nearest neighbor regression method described in Section 3.1, the following variables were selected as independent variables:

(1) LGBØVZ7 = LGB visibility at 0700 PST
(2) SANLAS7 = SAN - LAS ΔP at 0700 PST
(3) LAXIBC3 = 24-hour LAX inversion base height change at 14Z
(4) LAX9DIF = 24-hour LAX 950 mb temperature change at 14Z

The number of nearest neighbors used in the weighted average calculations were 5, 10, 15, and 20. For the linear weighted average calculations, the numbers were 15, 30, 45, 60, and 75. Root mean square errors for each of these methods are given in Table 3.6. As can be seen, the best results were obtained with 20 weighted average nearest neighbors, however, these results were not as good as the point-score method. For example, the nearest neighbor algorithm correctly predicted 88% of the episode conditions, and achieved a false alarm/probability of detection score of 28. The point-score method achieved 93% and 44, respectively.

RIVERSIDE - (RIVR)

Results of initial correlations between meteorological parameters and RIVR oxidant values were quite similar to those for Upland. Based on our experience in developing an algorithm for Upland, we pursued similar techniques for RIVR, using the LAX 850 mb and 950 mb temperatures and 24-hour temperature changes, as well as yesterday's oxidant persistence term.

Because of the tendency of regression analysis to underpredict high oxidant days, a successful working equation was not anticipated. However, such procedures were undertaken to verify the important meteorological parameters and to serve as a comparison for other algorithms. Results of the regression analysis are shown in Table 3.7. As in the case for UPLA, we were able to achieve a reasonably good fit (r = 0.82, N = 444) for the best regression equation, but high-end accuracy was not sufficient.

To develop a working algorithm, therefore, trends of the 850 mb and 950 mb temperatures and their 24-hour changes were examined for possible correlations to RIVR oxidant. As in the case of UPLA, the trend of LAX8TM4, alone was a reasonable predictor. As a result, the ensuing algorithm was keyed upon LAX8TM4.

Additional input from the 24-hour changes of LAX8TM4 and LAX9TM4 increased the predictive capabilities of the basic algorithm by adding sensitivity induced by changes in meteorology.

After testing the model against RIVRZMØ several alterations were made to improve prediction resolution and hence, prediction accuracy. These changes included a limit for the magnitude of the 24-hour change of LAX8TM4 and LAX9TM4. Also, upon examination of the model output, it was noted that for the month of October the algorithm had a tendency for

Table 3.6 Comparative Root Mean Square Errors (RMSE) of Various k Nearest Neighbor Estimators (La Habra)

	Estimator	RMSE (Values in PPHM)
1.	5-Weighted Average Nearest Neighbors	5.494
2.	10-Weighted Average Nearest Neighbors	5.186*
3.	15-Weighted Average Nearest Neighbors	5.099*
4.	20-Weighted Average Nearest Neighbors	5.077*
5.	15-Weighted Linear Nearest Neighbors	5.827
6.	30-Weighted Linear Nearest Neighbors	5.350
7.	45-Weighted Linear Nearest Neighbors	5.173*
8.	60-Weighted Linear Nearest Neighbors	5.132*
9.	75-Weighted Linear Nearest Neighbors	5.104*
10.	Linear	5.296

* Better than linear estimator

Table 3.7 Key Predictor Variables for Riverside Same-Day Regression

Variable	Coefficient
LAX8TM4	2.9
LAX9TM4	-3.4
SUMØPG7	-0.37
RIVRZMY	0.19
SDBØTM7	0.34
RBLTPH7	0.46
SAN5HC2	0.022
LGBØVZ7	0.13
CONSTANT	-2.09

Where: LAX8TM4 = LAX 850 mb Temp (°C) 14Z

LAX9TM4 = LAX 950 mb Temp (°C) 14Z

SUMØPG7 = |(LAX-DAG) + (SAN-LAS) + (SDB-VCV) + 6| Pressure gradient (mb) 07 PST

RIVRZMY = Yesterday's 1-Hr max OX at Riverside (pphm)

SDBØTM7 = SDB surface temp 07 PST

RBLTPH7 = (RBL - TPH) ΔP 07 PST

SAN5HC2 = SAN 500 mb height change (10 M) 12Z

LGBØVZ7 = LGB Surface visibility (mi) 07 PST overprediction. October tends to have increased offshore flow, which can result in high temperatures and low oxidant values (due to the westward push of pollutants). To adjust for the tendency to overpredict, persistence was added to the algorithm for the month of October.

Desirable characteristics of the final model included: (1) overall accuracy, (2) the ability to predict stage 2 episode levels, and (3) the ability to predict significant changes.

When tested against the dependent data set, the algorithm verified well against persistence in all categories,with special emphasis on the significant change days where 39% of the predictions were ± 2 pphm . (See Table 3.14.) Also, it should be noted that although the model failed to predict the site specific stage 2 episode, on the two days it did predict oxidant above 35 pphm, stage 2 concentrations did occur within the basin. For the 1977 verification similar improvements over persistence were evident with a lower mean absolute error and 40% of significant change day predictions within ± 2 pphm.

NEWHALL - (NEWH)

To obtain an initial calibration point for Newhall oxidant, the prediction algorithm determined for Riverside was applied to the Newhall oxidant data. Similarities in the two oxidant distributions existed so an attempt was made to tailor the Riverside prediction equation to the Newhall trends. For Newhall, oxidant concentrations exhibited greater persistence tendencies; therefore persistence (NEWHZMY) was included into the working algorithm. The resulting format for Newhall equation was similar to that for October - Riverside.

In the evaluation of the initial algorithm, it was observed that the model had a tendency to overpredict episode occurrences (i.e., it had a large false alarm rate). The model also resulted in a large mean absolute error with its cumulative prediction capabilities barely surpassing those of persistence. (It should be noted that persistence scored better in the rating format than most oxidant models.)

Several regression analyses were attempted to modify and improve the original algorithm.. First, regression analysis was focused upon the dependant variables in the primary equation. The resulting equation was fairly accurate with $R^2 = 0.51$ for 548 cases. When tested against the dependent data set the regression equation proved more accurate than the original algorithm tailored from that for Riverside. The regression equation predicted with a higher accuracy (85%) and a lower mean absolute error (3.3) than the original algorithm (see Table 3.15).

One critical feature of the model was its lack of sensitivity in predicting high oxidant concentrations. The structure of the regression acted to restrain fluctuating oxidant predictions. However, when the model did forecast episode levels it achieved a high degree of accuracy, with a minimal number of false alarms.

Further analyses were performed to expand the capabilities of both candidate algorithms. Upon inspection of various potential predictors, VGB5HC2 appeared to correlate well with the oxidant trends at Newhall. This correlation was not evident in subsequent regression analyses, where the inclusion of VBG5HC2 did little to significantly improve prediction capabilities.

Several additional regression attempts were performed upon selected variables with the end result of no significant improvement in predictive capabilities.

Using the initially created algorithm a correction term was constructed to modify the prediction resolution. The difference between predicted and observed oxidant was regressed against a series of independent variables to form the correction term however, upon implementation it failed to improve prediction accuracy. As a result, the regression equation developed from the original variables was used as the primary predictive algorithm.

3.3 ONE-DAY (24-HR) PREDICTIONS

Perhaps the most important prediction time period is the 24-hour prediction (issued at 3 PM). This issuance time is early enough to allow for the implementation of appropriate abatement strategies, yet late enough to incorporate important early-afternoon meteorological data (such as the mid-day LAX sounding). Also certain information about today's oxidant values (persistence terms) can be utilized. Using the results achieved in the development of the same-day algorithms, we sought to explore a variety of statistical methods to maximize the 24-hour capabilities. Final algorithms for the five key sites are given in the following subsection, with details of the chronological development given in Section 3.3.2.

3.3.1 Final Prediction Algorithms

(1) UPLAND

$$\begin{aligned} OX = {} & 0.36\,(SDB\emptyset TM3) - 1.34\,(LAXTPH3 - LAXTPH7) \\ & +0.33\,(PASDZM\emptyset) - 10.6 \end{aligned} \qquad (3.12)$$

where:

SDBØTM3 = Sandberg temperature at 1300 PST (°F)
LAXTPH3 = LAX-TPH ΔP at 1300 PST (mb)
LAXTPH7 = LAX-TPH ΔP at 0700 PST (mb)
PASDZMØ = Pasadena max hourly oxidant as of 1400 PST (pphm)

(2) RIVERSIDE - (AID & Regression - see Figure 3.12)

(3) NEWHALL - (AID & Regression - see Figure 3.13)

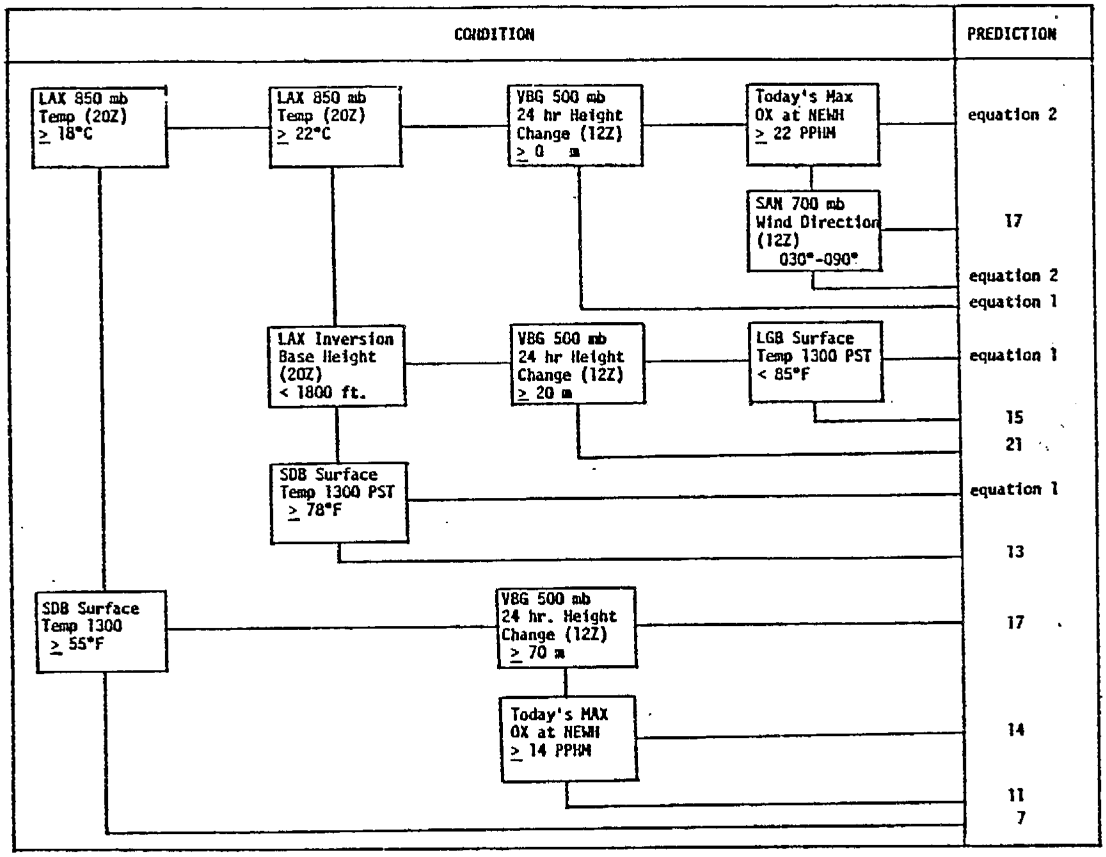
(4) DOWNTOWN L.A. - (AID - see Figure 3.10)

(5) LA HABRA - (AID - see Figure 3.14)

3.3.2 Chronological Development of Algorithms

UPLAND - (UPLA)

Of the five key sites, Upland is the most closely associated with the basin-maximum oxidant levels. Thus, the prediction of the oxidant levels at Upland is most critical for determining episode conditions for the following day. For these reasons, a considerable effort was undertaken along several statistical avenues.


CONDITION
PREDICTION
LAX 850 mb Temp (20Z) ≥ 18°C
LAX 850 mb Temp (20Z) ≥ 22°C
VBG 500 mb 24 hr Height Change (12Z) ≥ 0 m
Today's Max OX at NEWH ≥ 22 PPHM
equation 2
SAN 700 mb Wind Direction (12Z) 030°-090°
17
equation 2
equation 1
LAX Inversion Base Height (20Z) < 1800 ft.
VBG 500 mb 24 hr Height Change (12Z) ≥ 20 m
LGB Surface Temp 1300 PST < 85°F
equation 1
15
21
SDB Surface Temp 1300 PST ≥ 78°F
equation 1
13
SDB Surface Temp 1300 ≥ 55°F
VBG 500 mb 24 hr. Height Change (12Z) ≥ 70 m
17
Today's MAX OX at NEWH ≥ 14 PPHM
14
11
7

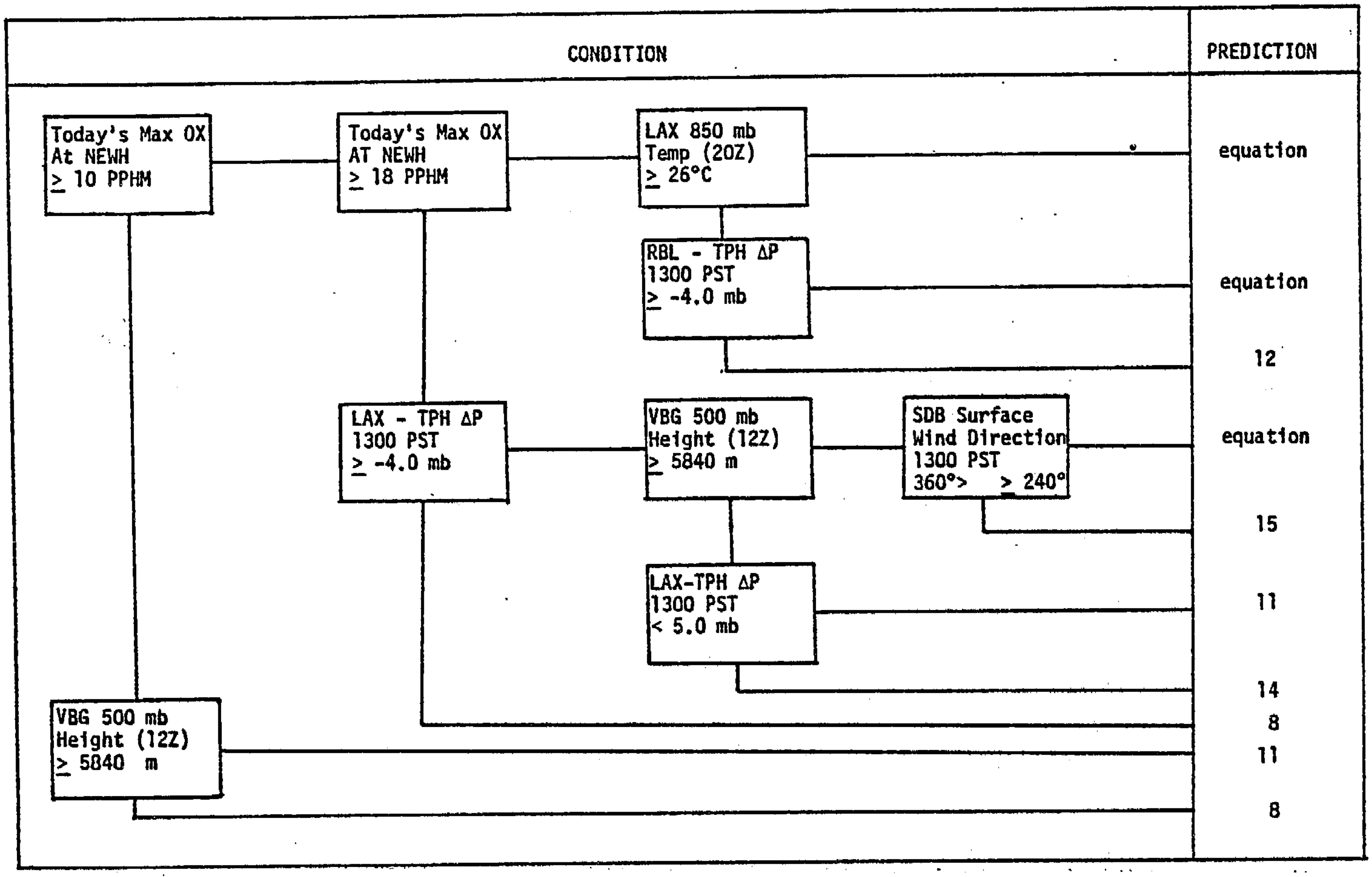
Equation 1 = -0.57 PLTOPC3 + 0.29 PASDZMØ + 0.16 SDBOTM3 - -0.55 LAXITT4
+0.72 LAX8TMØ -2.97

Equation 2 = -1.9 (SANLAS3 - SANLAS7) - 0.57 PLTOPC3 + 0.0044 LAXIBH4 + 0.91 LAX8TMØ
+0.90 (LAXITTØ - LAXITT4) + 0.57 LGBØVZ3 - 0.57 LAXTPH3 + 0.16 LGBØTM3 -17.78

24 Hour Riverside Oxidant Prediction Algorithm

Where: LAX8TMØ - 1PM 850 mb Temp at LAX (°C)
SDBØTM3 - 1PM Surface Temp at SDB (°F)
LGBØTM3 - 1PM Surface Temp at LGB (°F)
PLTOPC3 - Avg. of the 24 hr. Pressure Changes at WMC, RNO, TPH (mb)
PASDZMØ - Today's 1-hr. Max OX at PASD (PPHM)
LAXITT4 - 7AM LAX Inversion Top Temp (°C)
LAXITTØ - 1PM LAX Inversion Top Temp (°C)
SANLAS3 - 1PM Pressure Gradient Between SAN-LAS (mb)
SANLAS7 - 7AM Pressure Gradient Between SAN-LAS (mb)
LGBØVZ3 - 1PM Surface Visibility at LGB (miles)
LAXTPH3 - 1PM Pressure Gradient Between LAX-TPH (mb)

Figure 3.12 Decision-tree Prediction Algorithm for RIVR (one-day in advance). "YES" condition proceeds horizontally to the right; "NO" condition proceeds downward. Predicted values are in PPHM.


CONDITION
PREDICTION
Today's Max OX At NEWH ≥ 10 PPHM
Today's Max OX AT NEWH ≥ 18 PPHM
LAX 850 mb Temp (20Z) ≥ 26°C
equation
RBL - TPH ΔP 1300 PST ≥ -4.0 mb
equation
12
LAX - TPH ΔP 1300 PST ≥ -4.0 mb
VBG 500 mb Height (12Z) ≥ 5840 m
SDB Surface Wind Direction 1300 PST 360°> ≥ 240°
equation
15
LAX-TPH ΔP 1300 PST < 5.0 mb
11
14
8
VBG 500 mb Height (12Z) ≥ 5840 m
11
8

Prediction = 0.61 LAXØWV3 + 0.27 NEWHZMØ + 0.17 LAXIBTØ -2.0 DMTØTG3
-0.12 UPLAZMY + 0.23 LAXITTØ + 4.4

Where:

LAXØWV3 - 1PM Surface Wind Velocity At LAX (MPH)
NEWHZMØ - Today's 1-hour Max OX at NEWH (PPHM)
LAXIBTØ - 1PM LAX Inversion Base Temp (°C)
DMTØTG3 - 1PM Temp Gradient Between DAG-TRM (°C)
UPLAZMY - Yesterday's 1-hour MAX OX at UPLA (PPHM)
LAXITTØ - 1PM LAX Inversion Top Temp (°C)

Figure 3.13 Decision-tree Prediction Algorithm for NEWH (one-day in advance). "YES" condition proceeds horizontally to the right; "NO" condition proceeds downward. Predicted values are in PPHM. Equation is given on next page.

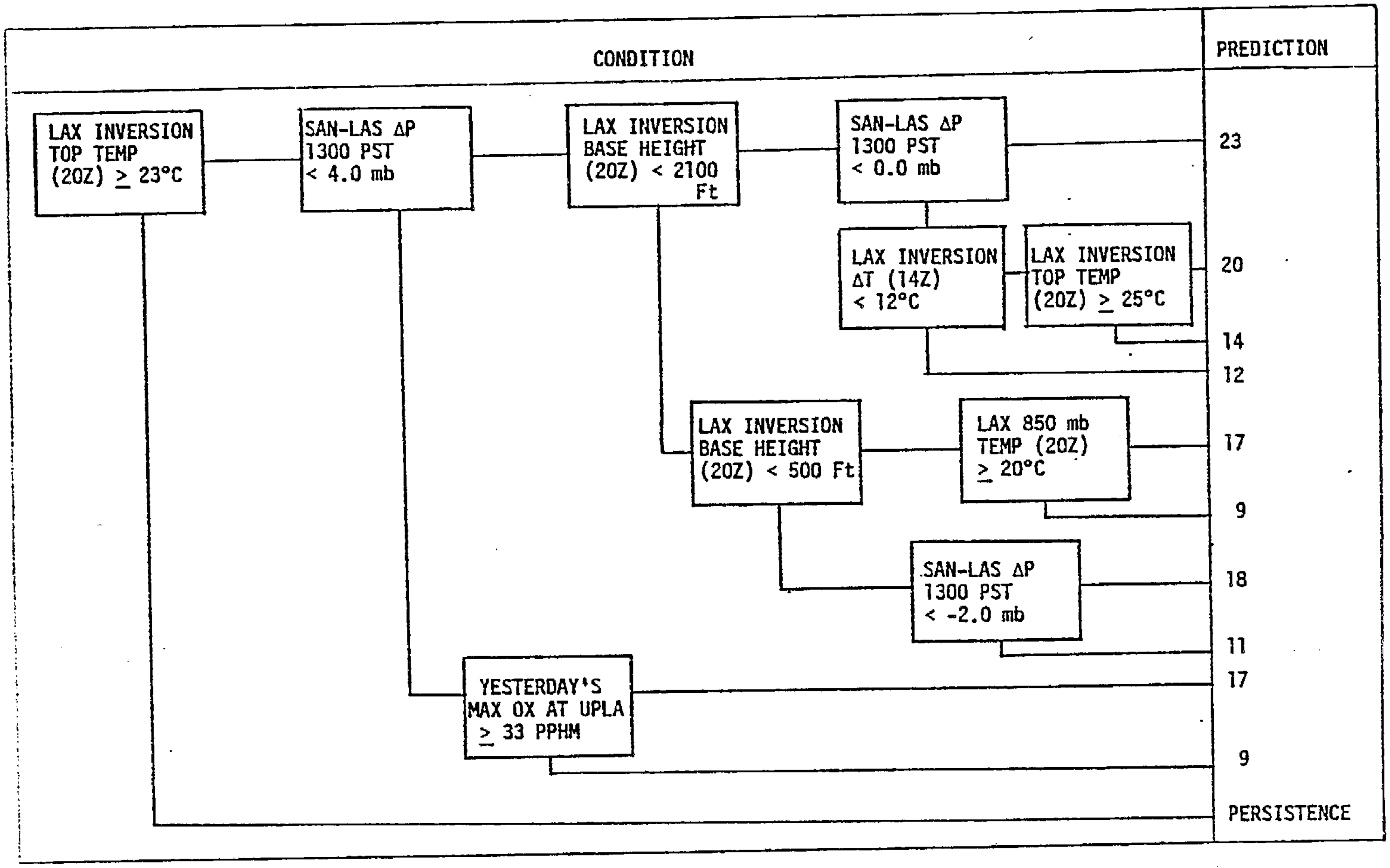

CONDITION
PREDICTION
LAX INVERSION TOP TEMP (20Z) ≥ 23°C
SAN-LAS ΔP 1300 PST < 4.0 mb
LAX INVERSION BASE HEIGHT (20Z) < 2100 Ft
SAN-LAS ΔP 1300 PST < 0.0 mb
23
LAX INVERSION ΔT (14Z) < 12°C
LAX INVERSION TOP TEMP (20Z) ≥ 25°C
20
14
12
LAX INVERSION BASE HEIGHT (20Z) < 500 Ft
LAX 850 mb TEMP (20Z) ≥ 20°C
17
9
SAN-LAS ΔP 1300 PST < -2.0 mb
18
11
YESTERDAY'S MAX OX AT UPLA ≥ 33 PPHM
17
9
PERSISTENCE

Figure 3.14 Decision-tree Prediction Algorithm for LAHB (one-day in advance). "YES" condition proceeds horizontally to the right; "NO" condition proceeds downward. Predicted values are in PPHM.

As would be expected, the relationship between meteorological variables and observed oxidant levels weakens as the prediction lead time increases. Initial screening regression yielded a prediction equation (which ultimately turned out to be the most effective 24-hour prediction algorithm) with a correlation coefficient (r) of 0.68. This compares to r = 0.83 for the best same-day equation.

In examining the selection of variables, we noted that the screening regression did not include specific pressure gradients which were thought to be important. Part of this effect could be attributed to the seasonal variations in the gradients (i.e. pressure gradients are more likely to be offshore in the August through October months than for the May through July period). Therefore, monthly "normal" values were computed for each of the parameters in the data base. For each day, a "departure from normal" was computed in terms of standard deviations from the mean. Regressions were run using "departure from normal" input variables. The results are summarized in Table 3.8. Even though pressure gradient terms were included in the resulting regression equations, the explained variance was not as great as the original equation.

Next, we applied the AID program to 20 potential predictors. The combination of conditions yielding the highest oxidant category (final node = 1) is defined by:

(1) LAX inversion top temperature (20Z) > 21°C
(2) LAX 850 mb temperature (20Z) > 22°C
(3) VBG 24-hr height change (12Z) > 0 m
(4) LGB visibility (1300 PST) > 5 miles
(5) RBL - TPH pressure gradient (1300 PST) > 2 mb

Neither the AID tree nor the regression equation successfully predicted any stage-2 events (equivalent to climatology), and the other aspects of the verification were very similar. We chose the regression equation over the AID output due to the increased simplicity of the regression (fewer input variables) and the continuous nature of regression output (as opposed to discrete prediction values from AID). However, we are including the full decision-tree for possible utilization, if desired.

Table 3.8 Key Predictor Variables for UPLA 24-Hour Regression

To Predict UPLA OX

Variable	Coefficient
LAX8TMØ*	0.38
SANWMC3*	-0.14
LAXIRHØ*	+0.10
LAXDAG3*	-0.33
LAX9TM4*	-0.16
LAXTRM3*	+0.17
Constant	+18.7

(N=488, r=0.56, S_e=7.4)

To Predict (UPLA OX)*

Variable	Coefficient
LAXITTØ*	0.17
SANWMC3*	-0.16
LAX9TM4*	-0.22
LAXIRHØ*	+0.13
LAXDAG3*	-0.40
LAXTRM3*	+0.22
LAX8TMØ*	+0.30
Constant	+0.1

(N=488, r=0.63, S_e=7.4)

Where:

LAX8TMØ = LAX 850 mb temperature at 19Z (°C)

SANWMC3 = SAN - WMC ΔP at 1300 PST (mb)

LAXIRHØ = LAX 1000 mb relative humidity at 19Z (%)

LAXDAG3 = LAX - DAG ΔP at 1300 PST (mb)

LAX9TM4 = LAX 950 mb temperature at 14Z (°C)

LAXTRM3 = LAX - TRM ΔP at 1300 PST (mb)

LAXITTØ = LAX inversion top temperature at (19Z)(°C)

* = departure from monthly mean (standard deviation)

An attempt was made to improve the regression results by using the residuals as the dependent variable in AID, with the full set of meteorological parameters as independent variables. Some key variables selected were SANLAS3, WEEKDAY, NEWHZMØ, and LAX8TMØ. However the overall improvement was only an additional 2% variance explained over the original regression. The added complexity of the combined algorithms was thought to be much greater in an operational sense than the small gain in accuracy, especially since the combined method still failed to predict any stage-2 conditions.

As a comparative method, we used the Box-Jenkins time series to predict UPLA maximum oxidant. As a brief review, the general Box-Jenkins model of order (p,d,q) is written in the form:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)\ \nabla^d z_t = (1 - \theta_1 B - \dots - \theta_q B^q) a_t$$

where B is the backshift operator $B^p_{z_t} = z_{t-p}$ and ∇^d is the difference operator $\nabla^d z_t = z_t - z_{t-d}$. If the transform is made from $\nabla^d z_t = z_{t-d}$ to w_t the model may be more simply written as a (p,q) of order d taking the form:

$$w_t - \phi_1 w_{t-1} - \dots \phi_p w_{t-p} = a_t - \theta_1 a_{t-1} - \dots \theta_q a_{t-q}$$

Here z_t is the observable variable (the initial time series), w_t is the differenced series and a_t is the white noise or random disturbances which cannot be predicted. Time series analysis, in the Box-Jenkins approach, may in fact be thought of as an attempt to reduce the residuals, or error terms, to uncorrelated noise.

For prediction purposes, we rewrite the model, with the appropriate differencing d (usually D or 1), giving w_t as:

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \dots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (3.13)$$

We know the past a's as the error between past forecasts and actual values. We take the current a, a_t, to be zero (0) as this is the statistical expectation of these uncorrelated terms. The problem of starting the forecasts can be most simply solved by taking all errors to be zero (0) until we have computed some actual errors. The first errors will of course not be accurate, but by doing this for the length of an observed series the errors shall converge to their uncorrelated state.

Explicitly, then, our forecasting model is

$$\hat{w}_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad (3.14)$$

replacing t with t+1 we can express our prediction for tomorrow in terms of yesterday and past days

$$\hat{w}_{t+1} = \phi_1 w_t + \cdots + \phi_p w_{t+1-p} - \theta_1 a_t - \cdots - \theta_q a_{t+1-q} \qquad (3.15)$$

where $\hat{w}$ is a forecast value and w is a known value.

For UPLA, we found that the model which predicted with the least error was of the form p=0, d=1, and q=3. For the equation:

$$Z_t = Z_{t-1} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} \qquad (3.16)$$

where Z = oxidant value
a = error terms
t = day
θ = best-fit coefficients

and the calculated θ values were determined to be:

$$\theta_1 = 0.29055 \pm 0.04029$$
$$\theta_2 = 0.38878 \pm 0.03884$$
$$\theta_3 = 0.17730 \pm 0.04036$$

Thus the resulting equation to predict tomorrow's oxidant level, Z_{t+1}, given today's oxidant level, Z_t, and the known error terms for the past 3 days, is:

$$Z_{t+1} = Z_t - 0.29a_t - 0.39a_{t-1} - 0.18a_{t-2}$$

Results of the daily predictions for the independent 1977 data set are shown in Table 3.11 which summarizes the verification scores. It can be seen that the time series method scored lower than persistence, but slightly better than the existing ARB objective equations. These results indicate that the time series approach (for directly predicting oxidant levels) is not an acceptable method.

Our next approach was to utilize the successful results from the same-day prediction algorithm. Since the only variables included in the UPLA same-day algorithm are LAX8TM4, LAX9TM4, and UPLAZMY, we sought methods to predict these variables which in turn could be used to predict UPLA oxidant. Prediction methods for the meteorological variables included linear regression, time series, and subjective techniques. However, since the data availability of the needed parameters are available earlier in the day (from morning observations), the items will be discussed in the next subsection on 30-hour prediction methods. Our final determination for the 24-hour prediction, therefore, indicates that the original regression equation is the most usable method, and has been shown to be more accurate than the existing ARB equation.

DOWNTOWN L.A. - (DOLA)

Initial regression techniques yielded the equation:

$$\begin{aligned} \text{DOLA OX} &= 0.34\ (\text{PASDZMØ}) + 0.14\ (\text{LGBØTM3}) \\ &\quad +0.15\ (\text{VBG5HC2}) - 4.7 \\ &\quad (N=501,\ r=0.60,\ S_e=4.2) \end{aligned} \tag{3.18}$$

where:

PASDZMØ = today's max oxidant at Pasadena

LGBØTM3 = LGB temperature at 1300 PST (°F)

VBG5HC2 = VBG 500 mb 24-hr height change at 12Z (10 m)

As in previous regression equations, the inability to predict high oxidant days ($\geq$ 20 pphm) was pronounced.

Using AID, a decision-tree was developed, shown earlier in Figure 3.10. Unlike other methods, the developed tree performed quite well, with an excellent ability to correctly classify episode days. Over the three-year period (1974-1976) this method correctly predicted 16 out of 31 possible episode days with no false alarm days. On our "false alarm rate - probability of detection" two-dimensional scoring system, the improvement over all existing methods is substantial. For overall score (see Table 3.12), the value of 208 is considerably greater than any existing method. In particular, note that 97% of all forecasts were correct episode conditions, and that 50% of all forecasts were ±2 pphm. As mentioned in the previous subsection, this method scored better than any same-day method. Hence once the prediction is made, no same-day updates are necessary.

LA HABRA - (LAHB)

As in the previous cases, the first analyses attempted were linear regression methods. Since the best fit using all data points explained only 19% of the variance (r=0.44), we selected only the high (≥ 18 pphm) and low (≤ 5 pphm) days to force a better fit. While the correlation increased (r=0.58) the standard error of estimate was also increased (from 6.0 to 7.7 pphm). The difficulty in using regression can be attributed to the unique set of meteorological conditions necessary to produce episodes at LAHB. A summary of the resulting equations is given in Table 3.9.

Since the effects at LAHB are in many ways similar to DOLA, it was anticipated that AID could identify the specific set of meteorological conditions conducive to episodes. The resulting tree was given in Figure 3.4. As expected the use of the decision-tree as a predictive method, especially in the identification of episode days, was considerably better than regression. Interestingly, the initial split for both LAHB and DOLA trees was for the inversion top temperature ≥ 23°C. More importantly, the algorithm correctly predicted 58% of all episode conditions for the dependent data set compared to 31% for persistence.

Table 3.9 Key Predictor Variables for LAHB 24-hour Regression

All Days		LAHB OX ≤ 5, ≥ 18	
Variable	Coefficient	Variable	Coefficient
LAXIBHØ	-0.0012	LAXIBHØ	-0.0024
SANLAS3C	-0.48	SANLAS3C	-0.75
LGBØVZ3	-0.20	LGBØVZ3	-0.22
Constant	+14.0	Constant	+17.1

where:

LAXIBHØ = LAX inversion base height at 20Z (FT)

SANLAS3C = SAN - LAS ΔP at 1300 PST today minus
SAN - LAS ΔP at 1300 PST yesterday (mb)

LGBØVZ3 = LGB visibility at 1300 PST

RIVERSIDE - (RIVR)

The initial technique that was attempted to determine a working algorithm for the 24-hour prediction was decision-tree analysis. Using selected optimal variables, the primary decision-tree achieved a total explained variance 48% for 475 summer cases. The initial split in the decision-tree was based upon LAX8TMØ. Other key variables included SDBØTM3, SAN7WD2, VBG5HC2, LGBØTM3, NEWHZMØ, and LAXIBHØ.

The model had the ability to forecast high values of ozone (up to 29 pphm) with a reasonable prediction resolution. With the input of wind direction data, the algorithm increased its sensitivity (directional sensitivity was not present in linear regression alone) and was able to predict stage-1 episodes more accurately. It did not have the ability to forecast stage-2 episodes. As a result, several alterations were attempted to improve the primary model.

In general, the basic framework of the initial algorithm was sound. The only areas where improvement was necessary were in the ability to forecast stage-2 episodes and in particular the loss of prediction resolution evident in the discrete forecasts. To adjust the model, several attempts were made to combine the output of the primary tree and capabilities of multiple regression.

To increase the sensitivity of the model a selected set of final nodes from the high end of the oxidant distribution were grouped together for a regression analysis. The idea of grouping high oxidant values together was to develop a continuous forecast method with the ability to forecast stage-2 episodes. Since the primary tree acted as a base screen for these high oxidant concentrations, a new set of meteorological predictors was chosen for the ensuing regression analysis. In a sense, we are trying to optimize the use of many potential predictors.

Two equations were produced from this analysis using different combinations of top nodes. The resulting equation proved to increase forecast resolution significantly. The algorithm did not forecast stage-2 concentrations for the dependent data set, yet given the right combination of parameters, a stage-2 prediction was possible. The only drawback of this analysis was in the increased complexity of the algorithm.

This procedure was repeated for middle class groupings of nodes to improve stage-1 forecasting. Several nodes were selected including both higher and lower class nodes, to expand the range of the prediction. The inclusion of this secondary equation allowed the model to catch several additional episodes (stage-1); however, it did overpredict for a larger percentage of oxidant values.

Several additional regression runs were made for RIVRZM1. They included a low class regression analysis (grouping bottom nodes together) and binary split regressions. For the lower class node regressions prediction capabilities were not improved sufficiently to warrant the increased complexity. The binary split analysis (using the initial split of the tree as a high-low indicator) also did not work well as a forecast algorithm.

One additional regression attempt was made--that relating a selected set of predictors to the total oxidant data set. The resulting RIVRZM1 equation is given below.

$$\begin{aligned} OX = {} & 0.063\ \text{LAX8TMØ} - 0.094\ (\text{LAXTPH3} - \text{LAXTPH7}) - 0.034\ \text{VBG5HC2} \\ & +0.17\ \text{SDBØTM3} - 0.39\ \text{LAXITT4} + 0.17\ \text{PASDZMØ} - 2.21 \end{aligned} \tag{3.19}$$

$(N=435,\ r=0.67,\ S_e=5.2)$

where:

LAX8TMØ = LAX 850 mb temperature at 20Z (°C)
LAXTPH3 = LAX - TPH ΔP at 1300 PST (mb)
LAXTPH7 = LAX - TPH ΔP at 0200 PST (mb)
VBG5HC2 = VBG 500 mb 24-hour height change at 12Z (10 m)
SDBØTM3 = SDB temperature at 1300 PST (°F)
LAXITT4 = LAX inversion top temperature at 20Z (°C)
PASDZMØ = Pasadena max hourly oxidant today (pphm)

NEWHALL - (NEWH)

With the same day Newhall analyses demonstrating the difficulty of predicting oxidant particularly about the stage-1 episode level, a different approach to the development of an accurate forecast algorithm was made. The initial procedure undertaken was to develop a decision-tree for 24-hour Newhall oxidant. The tree was developed from a series of highly correlating variables of whom NEWHZMØ was the key. Overall, the tree explained 46.9% of the variance.

Other variables that were important included VBG5HT2, SDBØWD3, UPLAZMY, LGBØVZ3, LAX8TMØ, and LAXTPH3. The first two major splits, though, depended upon NEWHZMØ. Of the final nodes produced, only one was able to forecast values of oxidant greater than 20 pphm. As a result, additional sensitivity was necessary for a working algorithm.

Stepwise regression, performed on selected terminal nodes (the highest nodes) used several additional predictors to increase stage-1 prediction accuracy. Various algorithms were attempted using different potential predictors. The most successful prediction equation increased the total variance explained to 55%. The ability to forecast episodes was increased due to a basic reduction in false alarms.

Several additional regression equations were attempted to better define NEWHZM1. Of these none improved the decision-tree predictions significantly. All additional regression analyses were performed using NEWHZM1 versus a determined set of optimal predictors. The data set was also regressed against the same day predictors; however, persistence becomes too dominant in the resulting equations, causing poor forecasting capabilities. The final day-in-advance Newhall regression equation that can be used as an alternate equation is essentially modified persistence:

$$\begin{aligned}\text{NEWHZM1} = {} & 0.50\ \text{NEWHZMØ} + 0.019\ \text{LAX8TM4} \\ & +0.013\ (\text{LAX9TM4} - \text{LAX9TMY}) + 3.16\end{aligned} \tag{3.20}$$

where:

LAX8TM4 = 14Z 850 mb temperature at LAX (0.1°C)

LAX9TM4 = 14Z 950 mb temperature at LAX (0.1°C)

LAX9TMY = 14Z 950 mb temperature yesterday at LAX (0.1°C)

(N=504, r=0.62, S_e=4.9)

3.4 THIRTY-HOUR PREDICTION ALGORITHMS

The 30-hour final prediction algorithms for the five key oxidant stations are based upon the "perfect prog" analysis. Each algorithm uses the input of both "in hand" morning meteorological variables and the output of selected variables forecast for the next day from the NWS numerical progs. Perfect prog analyses are based upon the premise that the numerical models produce valid estimates of same-day conditions permitting accurate oxidant prediction.

A variety of other methods were employed, however the statistical relationships between data in-hand this morning and observed values tomorrow afternoon (30-hour prediction) were satisfactory only under persistent conditions. This strongly suggests that the use of historical data for predicting oxidant concentrations 30 hours in advance is not appropriate, due to the dynamic meteorological changes that can occur.

To overcome these problems, some estimates of predicted meteorological conditions were necessary. As will be discussed in Section 3.4.2, two prediction features were evaluated: (1) subjective prediction of key meteorological variables, and (2) use of numerical progs (LFM 500 mb height data). Our results indicated that the numerical progs provided the most stable features, and in fact, achieved greater accuracy in some instances than the most successful 24-hour algorithms. The final 30-hour algorithms presented below, therefore reflect the "perfect prog" approach.

3.4.1 Final Forecast Algorithms

DOWNTOWN L.A. - (DOLA)

$$\text{OX} = 0.35\ (\text{VBG5HT2}_p) = 0.17\ (\text{SUMØPG7}) + 0.12\ (\text{DOLAZMY}) = 190.4 \qquad (3.21)$$

$$(N = 447 \quad r = 0.62 \quad r^2 = 0.38 \quad S_e = 4.2)$$

UPLAND - UPLA

$$\text{OX} = 0.68\ (\text{VBG5HT2}_p) - 0.29\ (\text{VMW5HT2}_p) + 0.15\ (\text{UPLAZMY}) + 0.25\ (\text{VBG5HC2}_p) - 375.7 \qquad (3.22)$$

$$(N = 442 \quad r = 0.69 \quad r^2 = 0.47 \quad S_e = 6.5)$$

RIVERSIDE - (RIVR)

$$\mathrm{OX} = 0.61\ (\mathrm{VBG5HT2}_p) - 0.23\ (\mathrm{VMW5HT2}_p) + 0.18\ (\mathrm{VBG5HC2}_p) - 336.0 \tag{3.23}$$

$$(N = 449 \qquad r = 0.66 \qquad r^2 = 0.44 \qquad S_e = 5.3)$$

LA HABRA - (LAHB)

$$\mathrm{OX} = 0.36\ (\mathrm{VBG5HT2}_p) - 0.12\ (\mathrm{VMW5HT2}_p) + 0.23\ (\mathrm{VBG5HC2}_p) + 0.14\ (\mathrm{LAHBZMY}) - 197.5 \tag{3.24}$$

$$(N = 408 \qquad r = 0.50 \qquad r^2 = 0.25 \qquad S_e = 5.7)$$

NEWHALL - (NEWH)

$$\mathrm{OX} = 0.26\ (\mathrm{VBG5HT2}_p) + 0.55(\mathrm{LAX8TM4}) + 0.0006\ (\mathrm{LAXIBH4}) - 146.0 \tag{3.25}$$

$$(N = 454 \qquad r = 0.53 \qquad r^2 = 0.28 \qquad S_e = 5.3)$$

where:

(units)		
(10 m)	$\mathrm{VBG5HT2}_p$	= VBG 500 mb height for 12Z tomorrow (from LFM progs)
(10 m)	$\mathrm{VMW5HT2}_p$	= 500 mb height difference between VBG and WMC for 12Z tomorrow (from LFM progs)
(10 m)	$\mathrm{VBG5HC2}_p$	= 500 mb 24-hour height change at VBG: tomorrow's 12Z height (from LFM progs) minus today's value
(mb)	SUMØPG7	= sum of the pressure gradients at 15Z today: $\lvert(\mathrm{SAN{-}LAS}) + (\mathrm{SBD{-}VCV}) + (\mathrm{LGB{-}DAG}) + 6\rvert$
(°C)	LAX8TM4	= LAX 850 mb temperature at 14Z today
(ft)	LAXIBH4	= LAX inversion base height at 14Z today
(pphm)	DOLAZMY	= yesterday's max oxidant at DOLA
(pphm)	UPLAZMY	= yesterday's max oxidant at UPLA

3.4.2 Development of the Final Forecast Algorithms

To produce a 30-hour forecast for oxidant at the five key stations several methodologies were attempted including: regression analysis, day in advance meteorological prediction and perfect prog regression analysis. The initial procedure incorporated regression analysis between meteorological variables and tomorrow's maximum oxidant at Upland (UPLAZM1). UPLAZM1 was chosen because of its frequency and severity of OX episodes (both stage 1 and stage 2). Selected combinations of meteorological variables available at the forecast period (MFR5HC2, LAX8TM4, SANLAS7, VBG5HT2, LAX8TMY) were regressed against UPLAZM1. Regressions were segregated by VBG5HT2 ($\geq$585 and <585) and LAX8DIF (LAX8TM4-LAX8TMY) $\geq$0 and <0, to improve prediction resolution. The concept of selecting discrete intervals of different predictors is analogous to setting a screen -- hopefully eliminating lower oxidant days from the full analysis.

The resulting regression equations proved to be unreliable as forecast algorithms. Shortcomings were evident in their inability to forecast stage-2 episode concentrations and in the low amount of variance explained ($R^2 = 0.37$ for the most productive equation). A large average standard error suggested the randomness of the predicted day-in-advance OX.

The major factors contributing to the poor correlations were the dynamic changes in meteorological conditions over the 30-hour period which were not predictable from initial conditions. The inability to predict the direction of a trend was particularly obvious from the resulting forecast. In effect, persistence was as good a predictor as the best linear fit.

To overcome this difficulty, we applied a two-step procedure: (1) predict key meteorological parameters on a 24-hour basis (predict tomorrow morning's condition based on this morning's data), and (2) input the results into the same-day algorithms. Three prediction methods were applied for the 24-hour meteorological prediction: (1) linear regression, (2) time series, and (3) subjective procedures.

Using linear regression, several key variables were predicted (LAXIBH4, LAX9TM4, and LAX8TM4). As in many of the previous regression analyses, significant changes were not caught until after the fact. Also the direction of change was not accurately forecast. The inclusion of persistence into the regression acted to stabilize the trends; however, rapid changes were suppressed in favor of the existing trend. In general, the regression equation proved to be inadequate for the designed purpose, again tending toward persistence.

It should be noted that, for UPLA, the same-day algorithm is quite sensitive to the direction of change of both the 850 mb and 950 mb temperatures. Thus a prediction of small temperature decreases, when in fact small temperature decreases occurred, could be substantially in error. Therefore the direction of change is important for the two-step prediction procedure to be accurate.

Time series analysis was then tested to obtain improved estimates of several key predictors. All of the time series models (LAX8TM4, VBG5HC2, LAXIBH4, PLTOPC7, SANLAS7, MFR5HC2) represented modified forms of persistence. In particular, the change variables VBG5HC2, PLTOPC7 and MFR5HC2 expressed little sensitivity to actual observed changes in the magnitude and direction of those variables. For all models major fluctuations in the existing variable distribution were suppressed. The output of the LAXIBH4 time series was extremely similar to that of the regression analysis where persistence played a dominant role. The results of the time series analysis failed to produce any valid met forecast for the ensuing day. The time series equations are given in Table 3.10.

The validity of subjective forecasting was also examined to determine whether or not it could be intermeshed with the objective forecasting system. Variables subjectively forecast daily included LAX8TM4, LAX9TM4, LAXIBH4, LAX1DT4, and SUMØPG7.

Prediction data were obtained from AQMD records for the 1974-1977 period. Since the AQMD 30-hour prediction is based on subjective meteorological input into an objective model, these prediction parameters were readily available. To obtain an initial "calibration" of the accuracy of the subjective predictions, the predicted values were regressed against the actual values. Correlation coefficients for all of the variables were above 0.70. However, from an examination of the residuals, predictions of significant change days were poor (both in magnitude and direction of the change). To determine the overall accuracy, we used the existing UPLA algorithm to estimate the oxidant persistence term, and the subjective predictions of the 850 mb and 950 mb temperatures. These were verified for both the 1974-1976 and the 1977 data bases. Table 3.11 gives the UPLA 1976 and 1977 verification scores. For comparative purposes, the subjective input prediction scored as follows:

Data Base	T_c -	10E +	T_2 +	C ÷	P_2 =	Rating
1977	68	61	30	10	25 =	72
1974-1976	68	62	27	12	28 =	73

Table 3.10 Summary of Time Series Prediction Equations for Key Meteorological Variables

Variable		Equation
LAXIBH4	=	$0.58\ y_t + 754$
LAX8TM4	=	$y_t - 0.21_{-t-1} - 0.39_{t-2} - 0.17_{-t-3}$
VBG5HC2	=	$0.52y_t - 0.44a_t - 0.27a_{t-1} - 0.17a_{t-2}$
SANLAS7	=	$y_t - 0.32a_t - 0.39a_{t-1} - 0.19a_{t-2}$
MFR5HC2	=	$y_t - 0.08a_t - 0.45a_{t-1} - 0.25a_{t-2}$
PLTOPC7	=	$0.25y_t - 0.45a_t - 0.46a_{t-1}$

where y_t = today's parameter value

$a_{(t-d)}$ = error term on day, (t-d), for t = today,
d = number of previous days

Note that these results are about equivalent to (though slightly less than) the current ARB subjective oxidant prediction capability. Thus there appears to be no advantage in subjectively predicting key meteorological parameters over direct subjective oxidant prediction.

As a forerunner to the "perfect prog" final algorithms, we sought to utilize upper air progs to predict the key meteorological variables. The premise was based upon a linear relationship between LAX8TM4 and LAX9TM4 versus 500 mg heights. By using this output as same day forecasts of alternate variables for the same period, the problem of forecasting weather by either purely statistical methods or subjective considerations could be avoided. Use of the output from the numerical models has the advantage of being closer (in time) to the oxidant condition.

Using VBG5HT2 and MFR5HC2, a regression equation was generated for LAX8TM4. The regression equation explained 67% of the variance in the 850 mb temperature yet had difficulty predicting significant changes, in particular the direction of the significant change. As previously mentioned, the directional change inaccuracy accounted for the lack of good prediction method.

Another method of using numerical progs involved scatterplot analysis. By plotting oxidant as a function of two variables obtained from the 500 mb progs, visual analysis of potential screens (thresholds of meteorological variables) defining high oxidant days was made. Several combinations were attempted including oxidant as a function of VBG5HC2 and SMO5HT2 (SAN5HT2-OAK5HT2) and oxidant as a function of VBG5HT2 and (VMM5HT2 + VMW5HT2)/2, (see Figures 3.15 and 3.16). In each scatterplot the darkened triangle represents a stage-2 episode. In Figure 3.16 there is no clear threshold for high oxidant levels.

Figure 3.15 illustrates a reasonable set of threshold values of met variables for UPLAZMØ, screening about 20% of the total number of days with 92.7% probability of no-episode conditions. The intent was to refine the unresolved cases with additional variables, such that eventually a

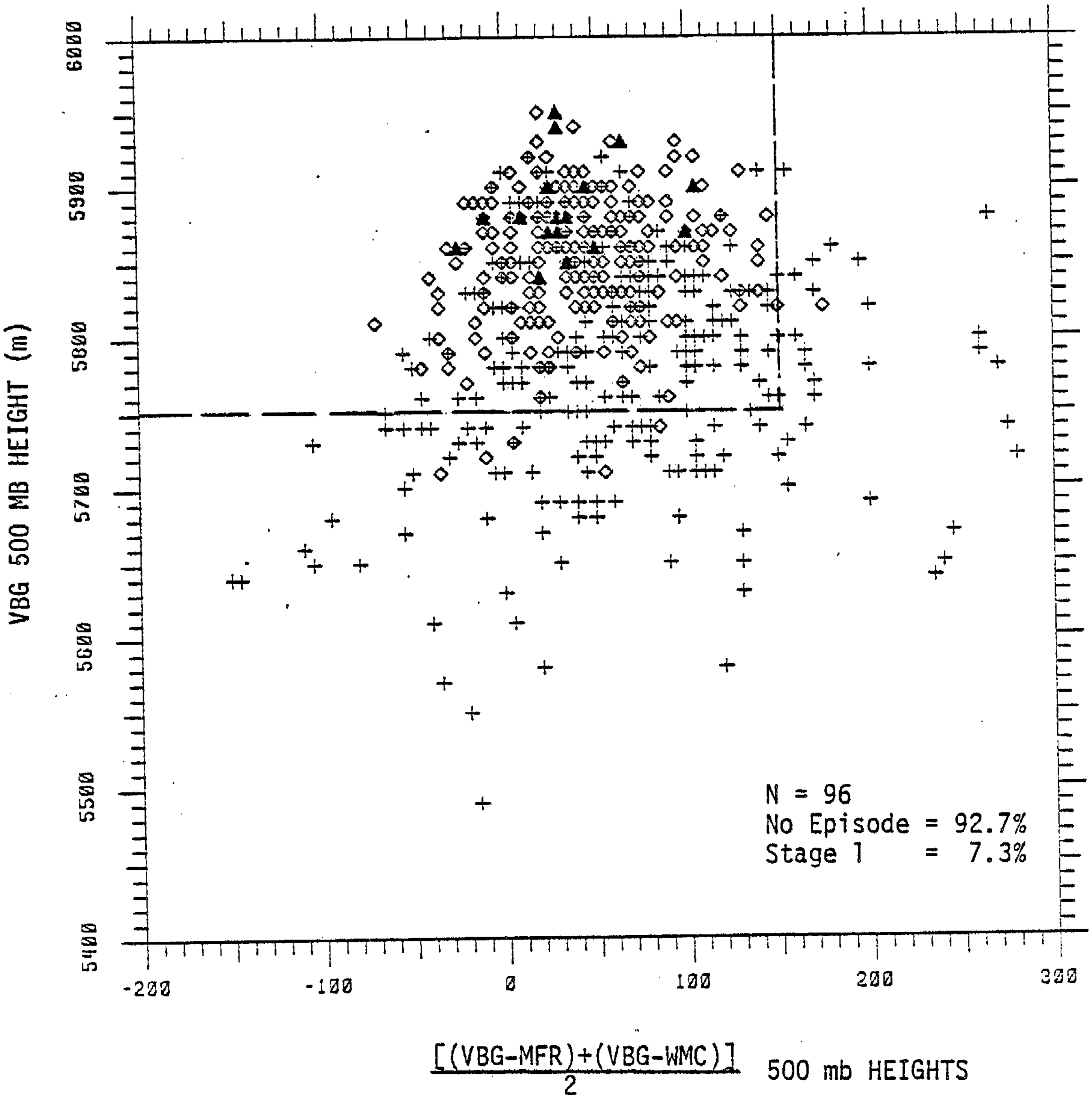


Figure 3.15 ;Upland Episode Conditions as a Function of Upper Air Data. (+ = No episode, ◇= Stage 1, ▲= Stage 2)

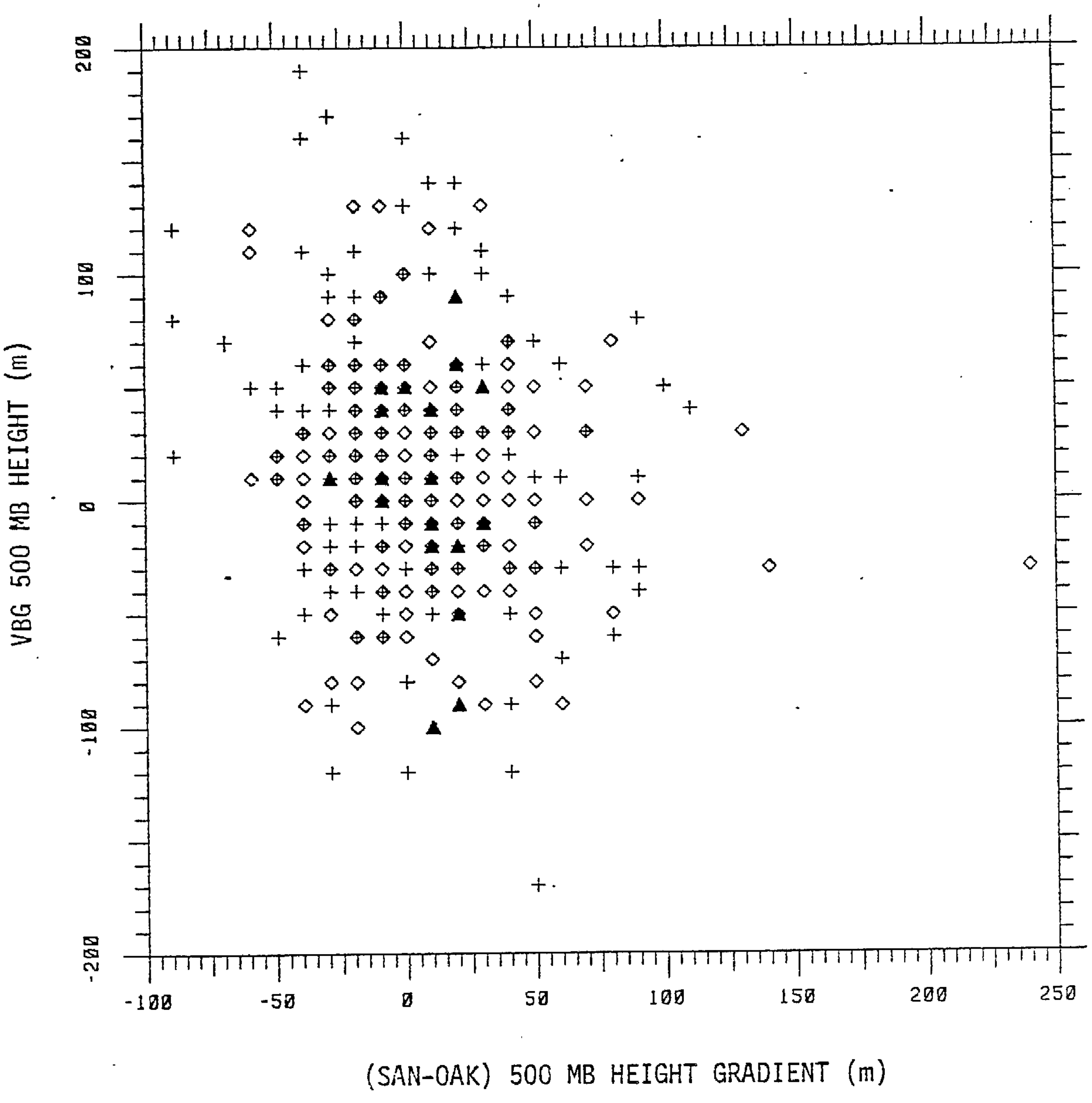


Figure 3.16 Upland Episode Conditions as a Function of Upper Air Data. No patterns are evident. (+ = No episode, ◇= Stage 1, and ▲= Stage 2)

combination of conditions would give a reasonable separation of stage 2 conditions. Ten successive screening scatterplots were generated (with input available from the numerical progs), but no clear-cut resolution was achieved.

As a result, we proceeded with the development of "perfect prog" regression equations, using observed 500 mb height data to achieve the greatest linear correlation. Upper air variables, including VBG5HT2, VBG5HT2-WMC5HT2 height differences and VBG5HT2 were combined with various in-hand variables including SUMØPG6, LAXIBH4, LAX8TM4 and UPLAZMY for the analysis. The resulting regression equation for Upland predicted with more accuracy than either set of equations determined by the initial regression or by the subjective/objective same day oxidant model. In fact, the verification using this method did better than the 24-hour algorithm! From Table 3.11, it can be seen that the perfect prog method scored 91, greater than any other day-in-advance prediction method. To test the reliability of actual "prog" input, key 500 mb height values were extracted from the 24-hour prediction panel of the NMC LFM prog package issued by 0900 PST for the May-October, 1977 period. These data provided a real-case test of the perfect prog method (since the developed equations were based on actual height data). The 1977 verification for Upland (Table 3.11) shows that the perfect prog equation scored 97, better than any other day-in-advance method, and was about equivalent to the current ARB subjective same-day score (98).

One additional method was attempted: to relate given meteorological variables to the change in daily oxidant values (ΔC). Thus if we could predict an expected ΔC, and we use the same-day oxidant prediction algorithms, we could predict tomorrow afternoon's oxidant. Using Upland oxidant data and 12 meteorological variables (LAX8TM4, MFR5HC2, VBG5HC2, VBG5HT2, LAXIBH4, LAXIDT4, OAK8TC2, LAX9TM4, LGBØTMY, SAN7RH2, SMM5HT2, and LAXITH4), the AID program was used to develop a decision-tree for ΔC. The results, however, yielded too much scatter, which when coupled with the same-day algorithms, proved to be substantially less accurate than the perfect prog method.

From the success of the perfect prog method at Upland, similar perfect prog equations were developed for the other four key sites. Verification results indicated that this method, as in the case for Upland, scored better than the 24-hour algorithms for both Riverside and La Habra. However, for DOLA and Newhall, the perfect prog method was not as good as the 24-hour algorithms, described in the previous subsection.

3.5 VERIFICATION

In Chapter 2 we described the method of evaluating predictive capabilities. Briefly stated, that method considers four key aspects of verification:

(1) episode prediction accuracy

(2) quantitative prediction accuracy

(3) significant change accuracy

(4) episode probability of detection/false alarm tradeoffs

Tables 3.11 through 3.15 summarize the verification results for both the dependent (1974-1976) and independent (1977) data sets at each of the five key sites. Results are presented according to methods available in a similar time period (i.e., all same-day prediction methods are grouped together and all day-in-advance methods are grouped together). The data compiled for the 1977 ARB and AQMD subjective predictions were taken from the final output records of these agencies. In other words, these represent the "official" forecasts issued.

It can be seen that in each case, the best method is one of the newly developed algorithms, and that similar results occurred on the independent data sets. Though none of the algorithms achieved the desired "goal" level established in the Phase I report, nevertheless, substantial improvement over all existing methods was achieved for most of the sites.

Using one-day persistence as a basis for comparisons, a ranking of all available methods on the independent data set (1977) is given in Table 3.16 The method used to construct the comparable scores is simply the fractional amount of the perfect score of the algorithm minus the equivalent value for one-day persistence. For example, the best method listed is for Newhall-same-day. From Table 3.15, the score of the new algorithm is .455. Subtracting the one-day persistence score of .165, we get an improvement over persistence (comparable score) of +.290.

As can be seen, most of the day-in-advance methods have negative scores (i.e., worse than one-day persistence). However, these predictions are usually made before today's value (persistence term) is known. Therefore, a comparison against one-day persistence may be too stringent a test of day-in-advance methods. (We used two-day persistence for the tabulations

Table 3.11 Overall Oxidant Prediction Rating for Upland

N	METHOD	T_c	-	10E	+	T_2	+	C	+	P (Pd,F/A)	=	R	R/PERFECT R
	PERFECT	100		0		100		100		100 (100,0)		400	
	DEPENDENT DATA SET: MAY-OCT 1974-1976												
	SAME-DAY PREDICTIONS												
454	1-DAY PERSISTENCE	67		57		28		0		25 (27,3.7)	=	63	.158
460	AQMD OBJECTIVE	74		45		35		24		26 (4, 1.0)	=	114	.285
142	ARB OBJECTIVE	54		74		28		18		20 (100, 6.3)	=	46	.115
452	ARB SUBJECTIVE	76		46		35		17		39 (31, 2.4)	=	121	.303
498	NEW ALGORITHM	75		46		35		18		45 (31, 1.9)	=	127	.318*
	ONE-DAY PREDICTIONS												
460	CLIMATOLOGY	59		67		22		17		25 (0,0)	=	56	.140
452	2-DAY PERSISTENCE	55		76		16		0		0 (5,4.9)	=	-5	-.013
434	AQMD OBJECTIVE	68		56		30		14		25 (9,1.8)	=	81	.203
460	AQMD SUBJECTIVE	68		57		28		12		34 (18, 1.6)	=	85	.213
133	ARB OBJECTIVE	58		64		30		10		25 (0,0)	=	59	.148
451	ARB SUBJECTIVE	68		54		29		17		21 (13, 3.0)	=	81	.203
515	NEW ALGORITHM (24-hr)	72		52		29		7		25 (0,0)	=	81	.203
507	PERFECT PROG (30-hr)	70		54		27		23		25 (0,0)	=	91	.228*
	INDEPENDENT DATA SET: MAY-OCT 1977												
	SAME-DAY PREDICTIONS												
175	1-DAY PERSISTENCE	74		50		38		0		20 (0, 1.1)	=	82	.205
176	TIME SERIES	72		55		33		0		18 (0, 1.7)	=	68	.170
172	ARB OBJECTIVE	68		59		24		12		0 (0, 5.1)	=	45	.113
172	ARB SUBJECTIVE	78		44		36		4		24 (0, 0.6)	=	98	.245
176	NEW ALGORITHM	85		36		49		20		70 (50, 0)	=	188	.470*
	ONE-DAY PREDICTIONS												
155	CLIMATOLOGY	57		66		23		4		25 (0,0)	=	43	.108
174	2-DAY PERSISTENCE	70		54		42		4		20 (0, 1.2)	=	82	.205
145	AQMD SUBJECTIVE	74		67		27		6		25 (0,0)	=	65	.162
151	ARB OBJECTIVE	64		55		29		0		25 (0,0)	=	63	.158
172	ARB SUBJECTIVE	70		52		32		0		24 (0, 0.6)	=	74	.185
177	PERFECT PROG (30-hr)	75		50		43		4		25 (0,0)	=	97	.243*

LEGEND

* = Best Method
N = Number of Predictions
T_c= Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes
Correct ±2 PPHM (%)

R = Rating
P = Score, Using Figure 2.4
P_d= Probability of Detection (%)
F/A= False Alarm Rate (%)

Table 3.12 Overall Oxidant Prediction Rating for Downtown Los Angeles

N	METHOD	T_c -	10E +	T_2 +	C +	P	(Pd,F/A)	=	R	R/PERFECT R
	PERFECT	100	0	100	100	100	(100,0)		400	
	DEPENDENT DATA SET: MAY-OCT 1974-1976									
	SAME-DAY PREDICTIONS									
546	1-DAY PERSISTENCE	90	35	47	0	12	(24,4.8)	=	114	.285
547	TIME SERIES	93	33	49	0	21	(0,1.3)	=	130	.325
302	ARB SUBJECTIVE	93	33	48	11	36	(24,2.0)	=	155	.388
	NEW ALGORITHM (24-hr)	(SEE BELOW)								
	ONE-DAY PREDICTIONS									
551	CLIMATOLOGY	94	39	40	30	25	(0,0)	=	150	.375
546	2-DAY PERSISTENCE	89	44	39	0	0	(12,5.5)	=	84	.210
547	TIME SERIES	94	38	41	11	25	(0,0)	=	133	.332
552	AQMD SUBJECTIVE	87	42	34	5	0	(44,9.9)	=	84	.210
300	ARB SUBJECTIVE	93	35	45	20	25	(12,2.3)	=	148	.370
493	NEW ALGORITHM (24-hr)	97	30	50	19	72	(52,0)	=	208	.520*
449	PERFECT PROG (30-hr)	93	35	41	8	25	(0,0)	=	132	.330
	INDEPENDENT DATA SET: MAY-OCT 1977									
	SAME-DAY PREDICTIONS									
180	1-DAY PERSISTENCE	97	31	57	0	20	(0,1.6)	=	143	.358
180	ARB SUBJECTIVE	98	34	47	0	52	(33,1.1)	=	163	.408
	NEW ALGORITHM (24-hr)	(SEE BELOW)								
	ONE-DAY PREDICTIONS									
178	2-DAY PERSISTENCE	97	38	40	20	20	(0,1.7)	=	139	.348
180	AQMD SUBJECTIVE	94	42	40	20	2	(0,3.9)	=	114	.285
180	ARB SUBJECTIVE	98	35	46	0	22	(0,0.5)	=	131	.328
182	NEW ALGORITHM (24-hr)	98	33	59	20	25	(0,0)	=	169	.423*

LEGEND

* = Best Method
N = Number of Predictions
T_c= Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes Correct ±2 PPHM (%)

R = Rating
P = Score, Using Figure 2.4
P_d= Probability of Detection (%)
F/A= False Alarm Rate (%)

Table 3.13 Overall Oxidant Prediction Rating for La Habra

N	METHOD	T_c	- 10E	+ T_2	+ C	+ P (Pd,F/A)	=	R	R/PERFECT R
	PERFECT	100	0	100	100	100 (100,0)		400	
	DEPENDENT DATA SET: MAY-OCT 1974-1976								
	SAME-DAY PREDICTIONS								
500	1-DAY PERSISTENCE	85	45	45	0	0 (27,7.4)	=	85	.213
480	NEW ALGORITHM	90	32	52	26	42 (28,2.0)	=	178	.445*
	ONE-DAY PREDICTIONS								
492	2-DAY PERSISTENCE	83	56	33	13	0 (15,8.3)	=	73	.183
486	NEW ALGORITHM (24-hr)	82	43	45	10	0 (56,13.9)	=	94	.235
468	PERFECT PROG	90	48	28	15	28 (4,0.2)	=	112	.280*
	INDEPENDENT DATA SET: MAY-OCT 1977								
	SAME-DAY PREDICTIONS								
180	1-DAY PERSISTENCE	96	27	58	0	44 (33,2.2)	=	171	.428
154	ARB SUBJECTIVE	95	29	56	0	52 (50,2.2)	=	174	.435
181	NEW ALGORITHM	96	23	65	43	48 (33,1.7)	=	229	.523*
	ONE-DAY PREDICTIONS								
181	2-DAY PERSISTENCE	93	35	50	0	8 (0,3.3)	=	116	.290
183	AQMD SUBJECTIVE	90	51	39	11	0 (33,7.7)	=	89	.223
153	ARB SUBJECTIVE	94	32	52	0	25 (50,2.2)	=	139	.348*
178	PERFECT PROG (30-hr)	96	43	33	29	23 (0,0.6)	=	138	.345

LEGEND

* = Best Method
N = Number of Predictions
T_c = Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes Correct ±2 PPHM (%)

R = Rating
P = Score, Using Figure 2.4
P_d = Probability of Detection (%)
F/A = False Alarm Rate (%)

Table 3.14 Overall Oxidant Prediction Rating for Riverside

N	METHOD	T_c	-	10E	+	T_2	+	C	+	P	(Pd,F/A)	=	R	R/PERFECT R
	PERFECT	100		0		100		100		100	(100,0)		400	
	DEPENDENT DATA SET: MAY-OCT 1974-1976													
	SAME-DAY PREDICTIONS													
540	1-DAY PERSISTENCE	71		45		33		0		23	(0,0.7)	=	82	.205
540	NEW ALGORITHM	78		42		39		39		25	(0,0.3)	=	139	.348*
	ONE-DAY PREDICTIONS													
539	2-DAY PERSISTENCE	64		62		25		4		23	(0,0.7)	=	54	.135
478	NEW ALGORITHM (24-hr)	74		42		36		12		25	(0,0)	=	105	.263
484	PERFECT PROG (30-hr)	73		43		38		20		25	(0,0)	=	113	.283*
	INDEPENDENT DATA SET: MAY-OCT 1977													
	SAME-DAY PREDICTIONS													
183	1-DAY PERSISTENCE	82		44		37		0		24	(0,0.5)	=	99	.248
184	NEW ALGORITHM	83		37		45		40		23	(0,1.1)	=	154	.385*
	ONE-DAY PREDICTIONS													
181	2-DAY PERSISTENCE	70		60		29		5		24	(0,0.6)	=	68	.170
181	AQMD SUBJECTIVE	75		49		36		11		25	(0,0)	=	98	.245
174	NEW ALGORITHM (24-hr)	74		48		31		18		24	(0,1.1)	=	99	.248
178	PERFECT PROG (30-hr)	74		45		40		12		25	(0,0)	=	106	.265*

LEGEND

* = Best Method
N = Number of Predictions
T_c= Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes Correct ±2 PPHM (%)

R = Rating
P = Score, Using Figure 2.4
P_d= Probability of Detection (%)
F/A= False Alarm Rate (%)

Table 3.15 Overall Oxidant Prediction Rating for Newhall

N	METHOD	T_c	-	10E	+	T_2	+	C	+	P (Pd,F/A)	=	R	R/PERFECT R
	PERFECT	100		0		100		100		100 (100,0)		400	
	DEPENDENT DATA SET: MAY-OCT 1974-1976												
	SAME-DAY PREDICTIONS												
546	1-DAY PERSISTENCE	82		39		42		0		0 (43,9.2)	=	85	.213
547	NEW ALGORITHM	85		33		48		3		60 (64.2.3)	=	163	.390*
	ONE-DAY PREDICTIONS												
545	2-DAY PERSISTENCE	75		56		30		0		0 (22,12.7)	=	49	.123
549	NEW ALGORITHM (24-hr)	84		34		47		3		35 (31,3.1)	=	135	.338*
480	PERFECT PROG (30-hr)	78		46		31		14		0 (38,13.3)	=	77	.193
	INDEPENDENT DATA SET: MAY-OCT 1977												
	SAME-DAY PREDICTIONS												
183	1-DAY PERSISTENCE	74		43		35		0		0 (59,13.1)	=	66	.165
182	NEW ALGORITHM	84		29		47		0		80 (88,1.6)	=	182	.455*
	ONE-DAY PREDICTIONS												
182	2-DAY PERSISTENCE	69		56		33		0		0 (53,15.4)	=	46	.115
182	AQMD SUBJECTIVE	74		45		37		7		0 (62,8.2)	=	73	.183
180	NEW ALGORITHM (24-hr)	79		42		41		0		0 (67,10.3)	=	78	.195*

LEGEND

* = Best Method
N = Number of Predictions
T_c= Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes Correct ±2 PPHM (%)

R = Rating
P = Score, Using Figure 2.4
P_d= Probability of Detection (%)
F/A= False Alarm Rate (%)

Table 3.16 Comparable Scoring and Ranking of Prediction Methods Using Independent Data Set (1977)

LOCATION	METHOD	PREDICTION TIME	COMPARABLE SCORE*	
Newhall	New Algorithm	Same Day	+.290	(+.177)**
Upland	New Algorithm	Same Day	+.265	(+.160)
Riverside	New Algorithm	Same Day	+.137	(+.143)
La Habra	New Algorithm	Same Day	+.095	(+.232)
DOLA	New Algorithm	24-hour	+.065	(+.235)
DOLA	ARB Subjective	Same Day	+.050	(+.103)
Upland	ARB Subjective	Same Day	+.040	(+.145)
Upland	Perfect Prog	30-hour	+.038	(+.070)
Newhall	New Algorithm	24-hour	+.030	(+.125)
Newhall	AQMD Subjective	30-hour	+.018	N/A
Riverside	Perfect Prog	30-hour	+.015	(+.078)
La Habra	ARB Subjective	Same Day	+.007	N/A
Riverside	New Algorithm	24-hour	±0.00	(+.058)
Riverside	AQMD Subjective	30-hour	-.005	N/A
Upland	ARB Subjective	24-hour	-.020	(+.045)
DOLA	ARB Subjective	24-hour	-.030	(+.085)
Upland	AQMD Subjective	30-hour	-.043	(+.055)
Upland	ARB Objective	24-hour	-.047	(-.010)
DOLA	AQMD Subjective	30-hour	-.073	(-.075)
La Habra	ARB Subjective	24-hour	-.080	N/A
La Habra	Perfect Prog	30-hour	-.083	(+.067)
Upland	ARB Objective	Same Day	-.092	(-.043)
La Habra	AQMD Subjective	30-hour	-.205	N/A

* Comparable Score = (Method Score - One Day Persistence Score).

**Comparable Score for 1974-76 dependent data set.

by site.) For purposes of establishing a fixed value by which we can compare all methods (same-day and day-in-advance), we decided to retain one-day persistence as the comparison term.

Another way of demonstrating the effectiveness of the final algorithms is to compare the results to the best existing prediction methods. These are given in Table 3.17. Note that the same day prediction methods did substantially better than the best existing methods, and while the improvement in day-in-advance capabilities is not as dramatic, there are appreciable results (except for La Habra). It should also be noted that in many instances, the best existing methods are subjective predictions. The new algorithms are completely objective, with input data readily available. Thus improvement over previously used objective systems is appreciable.

Table 3.17 Percent Improvement of New Algorithms Over Best Available Existing Methods (1977 Data)

LOCATION	BEST AVAILABLE EXISTING METHOD	PERCENT IMPROVEMENT*	
	SAME DAY		
UPLAND	ARB SUBJECTIVE	91.8%	(5.0%)**
DOLA	ARB SUBJECTIVE	3.7%	(34.0%)
LA HABRA	ARB SUBJECTIVE	20.2%	N/A
RIVERSIDE	PERSISTENCE	55.2%	(70.0%)
NEWHALL	PERSISTENCE	175.8%	(83.1%)
	DAY-IN-ADVANCE		
UPLAND	2-DAY PERSISTENCE	18.5%	(12.3%)
DOLA	2-DAY PERSISTENCE	21.6%	(38.7%)
LA HABRA	ARB SUBJECTIVE	-0.9%	N/A
RIVERSIDE	AQMD SUBJECTIVE	8.2%	N/A
NEWHALL	AQMD SUBJECTIVE	6.6%	N/A

*$\frac{\text{(NEW ALGORITHM SCORE - BEST METHOD SCORE)}}{\text{BEST METHOD SCORE}} \times 100.$

**COMPARISON TO DEPENDENT DATA SET.

3.6 EPISODE PROBABILITIES

In order to provide an estimation of the likelihood of an episode, given a predicted value, an analysis of the prediction error was performed. Empirical episode probabilities and confidence intervals were generated by processing the May through October observed data for the years 1974 - 76 with the output of the prediction algorithms developed previously. Statistics were calculated using two different methods, depending on the magnitude of the prediction. If the predicted value was less than 10 pphm, statistics were generated using the distribution of observed values. For values $\geq$ 10 pphm statistics were generated from the distribution of the differences between the logarithms of the observed and predicted values, or:

$$X = \ln \frac{Ca}{Cp}$$

Where X = error term
Ca = observed value
Cp = predicted value

The reason for using two procedures was to group the low-end predictions into one group (thus not having a discrete prediction value) and to prevent the errors of the low-end predictions from biasing the error distributions of the more significant predictions.

In the first case, a frequency distribution of observed values, sorted by increasing observed values, was computed, resulting in a cumulative probability distribution. Episode probabilities were easily derived from this distribution by first finding the probability that the observed value would be less than or equal to the specified episode value, and then determining the probability that the given value would be exceeded.

For predicted values $\geq$ 10 pphm, a distribution of the differences between the logs of the predicted and observed values was computed. This distribution was sorted as in the first case, and a cumulative probability distribution calculated. Episode probabilities and confidence intervals were calculated for each value of the predicted values from 10 pphm to 41 pphm. From this distribution, the probability of a difference less than or equal to an episode value was calculated. Episode probabilities were then generated as above, by taking one minus the interpolated cumulative probability. Two-tailed confidence intervals were computed using the same distribution, but instead of interpolating cumulative probabilities from some given difference value, difference values were interpolated from cumulative probabilities values. For example, for a confidence interval of 80%, differences corresponding to cumulative probabilities of 10% and 90% were calculated. These differences were then used to calculate low and high confidence intervals by determing the observed value which would lead to each difference value.

This procedure was repeated for each of the five key sites, and the remaining six sites on the ARB telemetry system (LENX, LONB, TEMC, RIVM. MTLE, and FONT.) Separate probabilities were also computed for each of the algorithms (same-day, 24-hour, and 30-hour perfect prog.) The complete set of probability tables is contained in Appendix C.

CHAPTER 4
SULFATE PREDICTION

4.1 BACKGROUND

Unlike gaseous pollutants which are continuously monitored, sulfates are collected over a 24-hour period and then analyzed in a laboratory. Thus it is not possible to "instantly" determine sulfate concentrations. Historically, the AQMD has sampled during a 24-hour calendar day period (midnight to midnight). This precludes rapid laboratory analysis. The ARB, on the other hand, samples from 10 a.m. to 10 a.m., so that the sample can be taken to the laboratory and analyzed within several hours. As indicated in Chapter 2, these results are important input for the ARB sulfate prediction equations. Since the sulfate value is recorded for the date in which the sample is removed, 14 hours (or 58%) of the total sample period occurs on the day preceding the indicated sample date. Only 10 hours (or 42%) occur on the same date. This is illustrated graphically in Figure 4.1, which shows the relationship between AQMD and ARB sampling times.

4.2 SELECTION OF KEY SITES

As described in Chapter 2, the ARB has developed regression equations for six SCAB locations: Anaheim, Azusa, Reseda, Riverside, Temple City, and Upland. Since these are the only sites where daily sulfate values are measured, these will represent the six key SCAB sites. Each of these sites is near an existing AQMD site, except Temple City, which is somewhat farther away from the AQMD Pasadena site. Statistics were generated comparing ARB to AQMD samples (see Table 4.1). Means and standard deviations were computed for each site, and for AQMD, ARBSAME, and ARBNEXT samples. To determine the differences between the samples, a weighted average of the ARB samples was obtained based on the proportionality of overlapping sampling time with the AQMD samples. It can be seen that the sample differences are very small (±5%) for Anaheim, Azusa, Reseda, and Riverside, while Pasadena-Temple City (-18.3%) and Upland (+45.1%) appear to be large. A t-test was performed to determine the

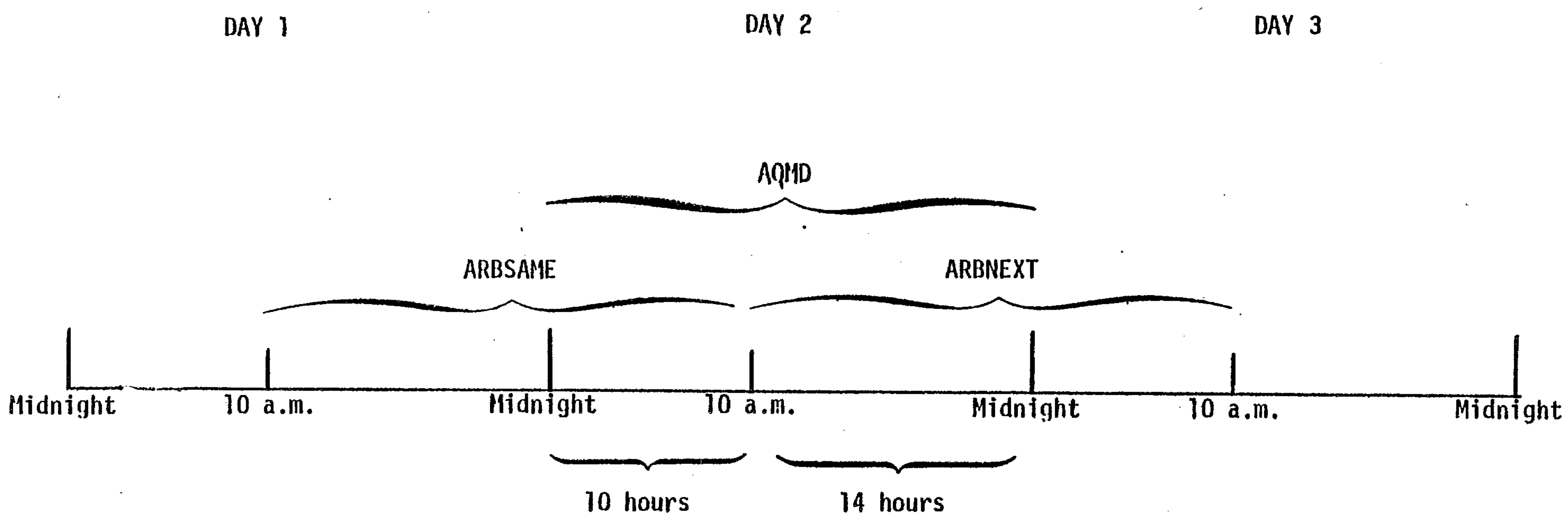


Figure 4.1 Sulfate Sampling Time Relationship between AQMD and ARB Samples. (ARBSAME is the recorded sample date equivalent to the AQMD sample date; ARBNEXT is the recorded sample date for the day after the AQMD sample date.)

Table 4.1 Sulfate Statistics Comparing Midnight (AQMD) to Midday (ARB) Samples*

	Number Of Samples	SAMPLE MEAN ± σ			(d) ARB WEIGHTED MEAN	% DIFFERENCE ($\frac{d - a}{a}$ x 100)
		(a) AQMD	(b) ARBNEXT	(c) ARBSAME		
ANAHEIM	16	11.8 ± 5.7	11.8 ± 5.8	11.2 ± 5.6	11.6	-1.7%
AZUSA	16	17.4 ± 8.0	18.4 ± 10.0	15.4 ± 7.7	17.2	-1.1%
RESEDA	14	12.6 ± 5.9	15.0 ± 6.6	12.6 ± 5.8	14.0	+2.9%
PASADENA (TEMPLE CITY)	56	14.2 ± 9.0	11.6 ± 8.7	11.5 ± 7.3	11.6	-18.3%
UPLAND	25	9.1 ± 4.5	13.0 ± 9.1	13.5 ± 8.1	13.2	+45.1%
RIVERSIDE	44	11.4 ± 7.3	12.0 ± 6.9	11.8 ± 6.9	11.9	+4.4%

*Statistics computed only for cases when both ARBNEXT and ARBSAME data were available for comparable AQMD sample (May-October, 1976-1977). In some instances sample sizes are small due to intermittent nature of both AQMD and ARB sampling.

significance of these differences. From these analyses (shown in Table 4.2), it can be seen that the Temple City-Pasadena differences are indeed significant, probably due to the distance between these sites. For Upland, the difference between AQMD vs ARBSAME is only marginally significant. (The lack of significance at Upland compared to Pasadena-Temple City is due to a smaller sample size for Upland.) It is difficult to explain the phenomenon at Upland, since monitoring locations are within two miles of each other.

It should also be noted that one outlier was eliminated from the data bases for Upland and Riverside. This occurred on September 9, 1976, for both locations. The following list illustrates the sulfate values reported on that date:

Upland:	AQMD	30.6	(ug/m^3)
	ARBSAME	6.8	"
	ARBNEXT	1.4	"
Riverside:	AQMD	44.3	"
	ARBSAME	1.5	"
	ARBNEXT	4.9	"
Fontana:	AQMD	7.9	"
Chino:	AQMD	6.8	"
San Bernardino:	AQMD	14.0	"
Temple City:	ARBSAME	7.9	"
	ARBNEXT	2.5	"

There are no obvious reasons for the inexplicably high values at Upland and Riverside AQMD samples. A check with the AQMD Eastern Zone did not reveal any additional information (e.g. the sample was not collected for more than 24-hours), nor did an examination of meteorological factors reveal any unusual events. Hence, these data were removed as outliers. (Scatter plots are given in Figures 4.2 and 4.3 which show the extent of the outliers.)

With the outliers removed, equations for the AQMD sampling times were generated using regression techniques. These are summarized in Table 4.3.

TABLE 4.2 T-Test Results to Determine Significance of the Differences of the Sulfate Means (ARB vs. AQMD)

STATION	COMPARED SAMPLES	DIFFERENCE (MEAN)	NUMBER OF CASES	t-VALUE	p-VALUE*	SIGNIFICANT
ANAHEIM	AQMD vs. ARBSAME	0.02	22	-0.02	0.98	NO
	AQMD vs. ARBNEXT	0.16	18	-0.23	0.82	NO
AZUSA	AQMD vs. ARBSAME	0.03	21	0.02	0.98	NO
	AQMD vs. ARBNEXT	-1.53	19	-1.54	0.14	NO
RESEDA	AQMD vs. ARBSAME	0.33	18	0.47	0.64	NO
	AQMD vs. ARBNEXT	-1.17	19	-1.60	0.13	NO
RIVERSIDE	AQMD vs. ARBSAME	0.51	45	0.43	0.67	NO
	AQMD vs. ARBNEXT	0.21	46	0.19	0.85	NO
TEMPLE CITY	AQMD vs. ARBSAME	2.69	59	4.90	<0.01	YES
	AQMD vs. ARBNEXT	2.64	59	6.56	<0.01	YES
UPLAND	AQMD vs. ARBSAME	-2.87	28	-2.03	0.05	MARGINAL
	AQMD vs. ARBNEXT	-1.79	30	-1.03	0.31	NO

* Significant at $p < .05$

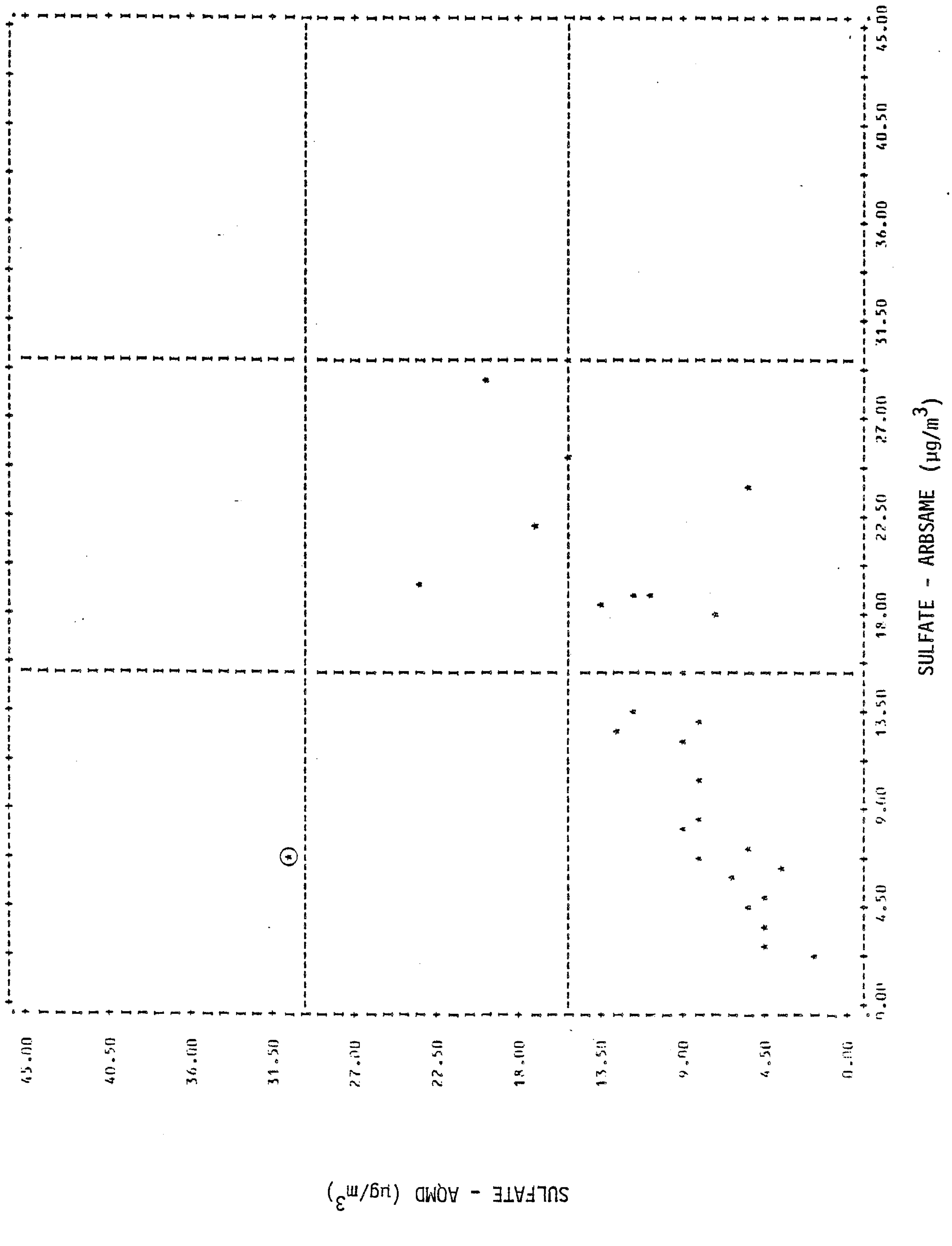


Figure 4.2 Scatterplot of AQMD vs. ARBSAME for Upland (Circled point is outlier)

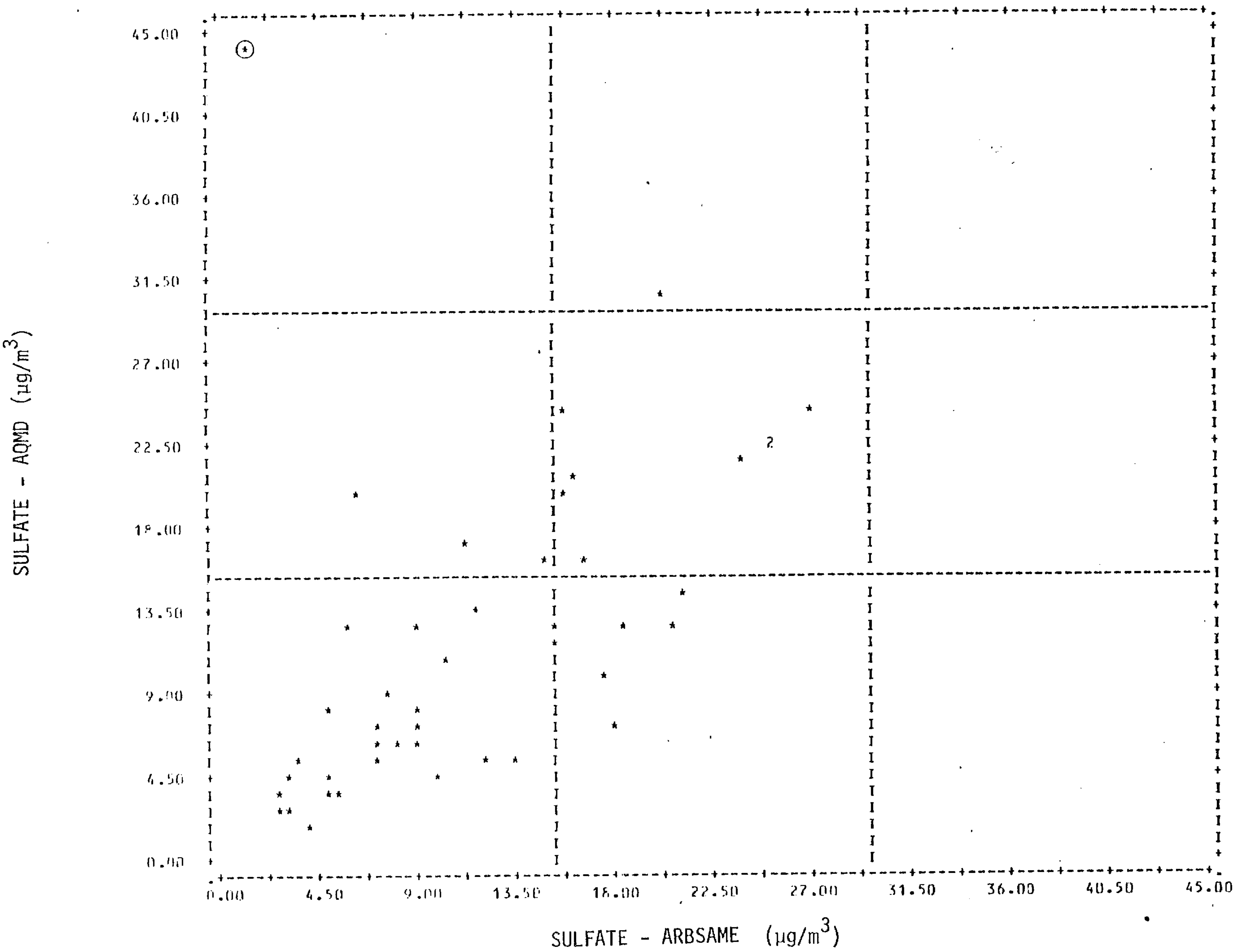


Figure 4.3 Scatterplot of AQMD vs. ARBSAME for Riverside (Circled point is outlier)

Table 4.3 Predictions of AQMD Samples from ARB Samples

STATION	EQUATION	N	R	R^2	SE
ANAHEIM	AQMD = 0.59 ARBNEXT + 0.44 ARBSAME - 0.1	16	0.93	0.87	2.2
AZUSA	AQMD = 0.53 ARBNEXT + 0.40 ARBSAME + 1.4	16	0.98	0.97	1.6
RESEDA	AQMD = 0.52 ARBNEXT + 0.43 ARBSAME + 0.3	14	0.95	0.91	1.9
PASADENA (TEMPLE CITY)	AQMD = 0.67 ARBNEXT + 0.45 ARBSAME + 1.2	56	0.97	0.93	2.4
UPLAND*	AQMD = 0.31 ARBNEXT + 0.16 ARBSAME + 2.9	25	0.87	0.77	2.3
RIVERSIDE*	AQMD = 0.65 ARBNEXT + 0.25 ARBSAME + 0.6	44	0.83	0.69	4.2

*OUTLIER REMOVED

N = Number of samples
R = Correlation Coefficient
R^2 = Variance explained
S_E = Standard error of the regression

Summary statistics are also shown. It is interesting to note that the equation coefficients for most sites are in the approximate proportionality of the sample overlap. Thus, given today's ARB sample value (persistence), and the ARB predicted value for tomorrow, the AQMD sample can be predicted accordingly.

4.3 METHODOLOGY

Due to the excellent sulfate verification results using the ARB equations (see Chapter 2), it was apparent that the most important improvement would not be to regenerate new sets of equations, but rather to diagnose the existing equations for weaknesses, and to modify them appropriately.

An analysis of the prediction errors was conducted by initially sorting the magnitude of the error by the persistence term (and noting the date). This allowed us to analyze the nature of the error and relate the overprediction/underprediction to meteorological variables. For example, moderate underpredictions at the low end (i.e. <15 ug/m^3) were not as important as moderate underpredictions elsewhere (i.e. $\geq$15 ug/m^3), in which episode conditions would not be correctly predicted. In particular, this procedure allowed us to identify conditions in which the prediction equation for Upland would "blow up" (i.e. predicted values would get very high).

On a case-by-case basis, we found that the most significant errors occurred under conditions with large inversion ΔT's (the change in temperature from the inversion base to the inversion top). Modifications to the equations for Upland, Riverside, and Reseda were constructed based, in large part, on the effects of the ΔT. Table 4.4 summarizes these results.

For Riverside, modifications are only necessary for ΔT $\geq$ 10 (the critical value necessary to implement the modification). In these cases, the predicted value is increased 1 μg/m^3 for every degree the temperature is greater than 8°C (the adjustment value). In the same manner, for Reseda, the critical ΔT value is 8°C, and adjustment value is 10°C. Note that (for Reseda) it is possible to reduce the value of the original predictions by as much as 2 μg/m^3. Also, in order to reduce underpredictions under certain inversion

Table 4.4 Modifications to the ARB Sulfate Equations

1. RIVERSIDE: $\mathrm{RIVR}_{\mathrm{MOD}} = \mathrm{RIVR}_{\mathrm{EX}} + \kappa(\Delta T - 8.0)$

$$\text{where } \kappa = \begin{cases} 0 \text{ if } \Delta T < 10 \\ 1 \text{ if } \Delta T \geq 10 \end{cases}$$

2. RESEDA: $\mathrm{RESD}_{\mathrm{MOD}} = \mathrm{RESD}_{\mathrm{EX}} + \kappa(\Delta T - 10.0)$

$$\text{where } \kappa = \begin{cases} 0 \text{ if } \Delta T < 8 \\ 1 \text{ if } \Delta T \geq 8 \end{cases}$$

NOTE: IF INV. BASE: $14 \leq \mathrm{IB} \leq 20$

THEN USE MAX $\left\{\begin{matrix}\mathrm{RESD} \\ \mathrm{AZUS}\end{matrix}\right\}$ CONTINUITY

3. UPLAND: $\mathrm{UPLA}_{\mathrm{MOD}} = \kappa[\mathrm{RIVR}_{\mathrm{MOD}} + (\Delta T - 6.0)] + \ell \cdot m(\mathrm{UPLA}_{\mathrm{EX}})$

$$+ \ell \cdot \eta\ (\mathrm{AZUS}_{\mathrm{EX}} - 1.0)$$

$$\text{where } \kappa = \begin{cases} 0 \text{ if } \Delta T < 10 \\ 1 \text{ if } \Delta T \geq 10 \end{cases}$$

$$\ell = \begin{cases} 1 \text{ if } \Delta T < 10 \\ 0 \text{ if } \Delta T \geq 10 \end{cases}$$

$$m = \begin{cases} 0 \text{ if } (\mathrm{AZUS}_{\mathrm{EX}} - \mathrm{UPLA}_{\mathrm{EX}}) < 0 \\ 1 \text{ if } (\mathrm{AZUS}_{\mathrm{EX}} - \mathrm{UPLA}_{\mathrm{EX}}) \geq 0 \end{cases}$$

$$\eta = \begin{cases} 1 \text{ if } (\mathrm{AZUS}_{\mathrm{EX}} - \mathrm{UPLA}_{\mathrm{EX}}) < 0 \\ 0 \text{ if } (\mathrm{AZUS}_{\mathrm{EX}} - \mathrm{UPLA}_{\mathrm{EX}}) \geq 0 \end{cases}$$

conditions, it is necessary to substitute Azusa persistence in the original Reseda equation.

Results for Upland were more complex, primarily due to the need to prevent the "blow up" feature of the existing equation. There are basically three terms in the modified equation. The first term, like the adjustments for Riverside and Reseda, are based on a critical $\Delta T \geq 10°C$. In this case, however, the adjustment is made for the Riverside modified equation. Since the Riverside equation has Upland continuity as one of the input parameters, it is physically sensible that the best results were achieved with this modification. In essence, it allows the predicted Upland value to be based on Upland persistence. The second term (for conditions $\Delta T \leq 10°C$ and Azusa > Upland) uses the existing Upland equation. The third term prevents the "blow up" condition by not allowing the predicted Upland value to exceed the predicted Azusa value.

4.4 VERIFICATION

Using the verification techniques described in Chapter 2, method scores were compiled for each of the six sulfate prediction sites. These are given in Table 4.5. Also shown, in Figure 4.4, are the probability of detection/false alarm rate scores (P-Scores) for each of the sites, with directional improvement over one-day persistence indicated. Note that for Temple City, Anaheim, and Azusa, the improvement by the original ARB equations was very good; hence, modification efforts to improve these algorithms were not able to substantially improve these predictions. For Reseda, Upland, and Riverside, the ARB equations provided only marginal improvement over persistence. The addition of the modifications substantially improved the prediction capability for each of these sites. The greatest improvement occurred in the P-scores and a better ability to predict within 2 $\mu g/m^3$ of the actual value.

A comparison to one-day persistence is given in Table 4.6. For each site, the improvement over persistence is at least +0.2, which is at least as good as every prediction algorithm for oxidant, except Upland and Newhall (same-day).

Table 4.5 Overall Sulfate Prediction Rating
(May-Oct 1977)

Method	T_c	-	10E	+	T_2	+	P	=	Rating	Rating ÷ Perfect Rating
PERFECT	100		0		100		100		300	
ANAHEIM										
Persistence	94		61		13		43		89	.297
ARB Equation	97		41		38		62		156	.520*
AZUSA										
Persistence	91		66		25		40		90	.300
ARB Equation	96		44		38		84		174	.580*
RESEDA										
Persistence	93		54		20		30		85	.283
ARB Equation	96		38		28		20		106	.353
Modified Equation	98		34		45		70		179	.597*
RIVERSIDE										
Persistence	93		50		28		44		115	.383
ARB Equation	92		34		32		43		133	.443
Modified Equation	95		30		46		65		176	.587*
TEMPLE CITY										
Persistence	93		64		14		46		89	.297
ARB Equation	96		43		33		67		153	.510*
UPLAND										
Persistence	92		49		25		40		108	.360
ARB Equation	95		39		34		40		130	.433
Modified Equation	96		36		42		71		173	.577*

*Best Method

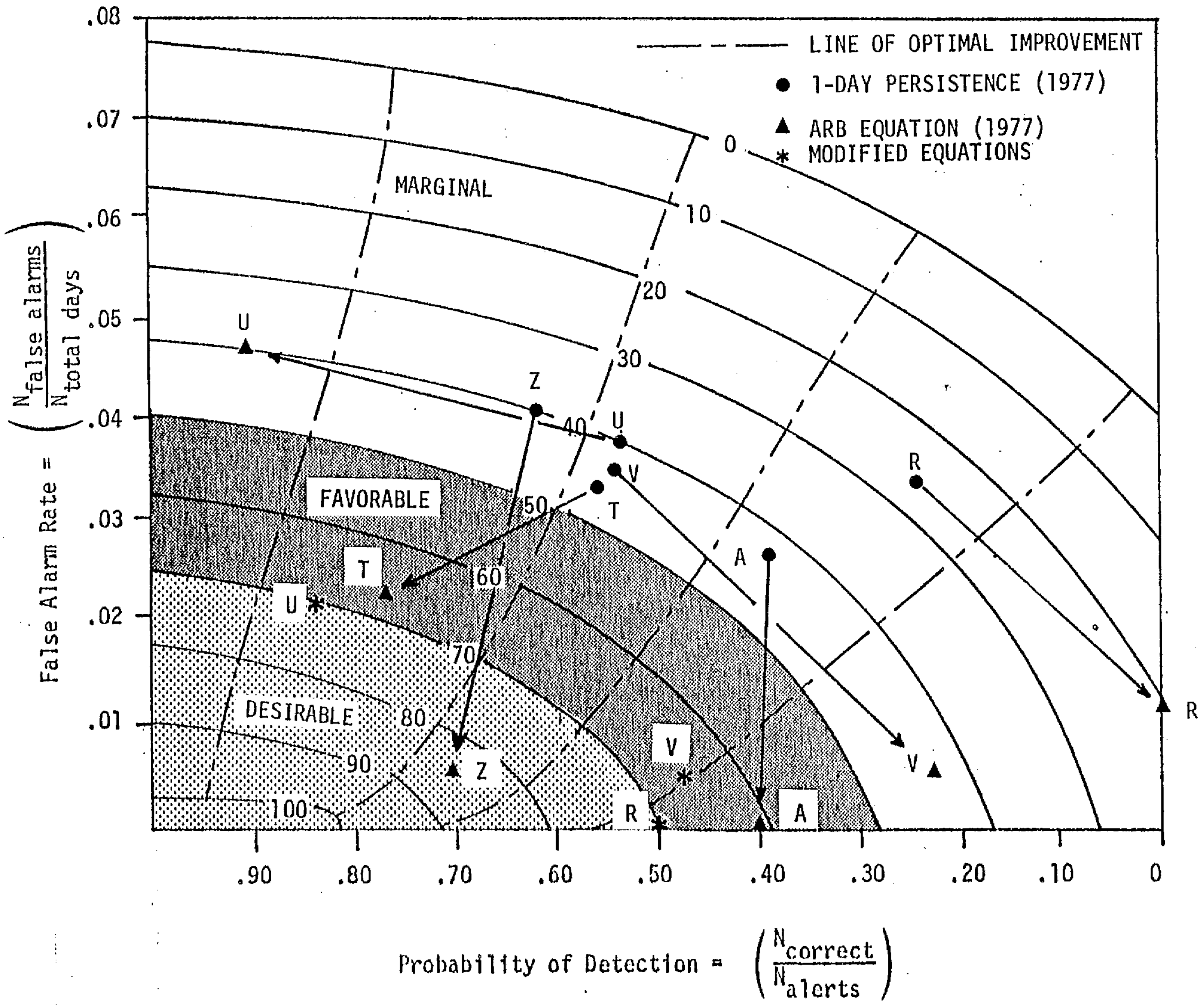


Figure 4.4 Scoring system evaluation for 1977 sulfate prediction. Letters refer to city (U = Upland, Z = Azusa, R = Reseda, A = Anaheim, T = Temple City, V = Riverside). Arrows show direction of improvement (degradation) over persistence.

Table 4.6 Comparable Prediction Scores
For Each Method versus One-Day Persistence

Location	Perfect Rating (PR)	Method Rating / PR	-	1-Day Persistence / PR	=	Comparable Ability
ANAHEIM (1)	300	.520	-	.297	=	+.223
AZUSA (1)	300	.580	-	.300	=	+.280
RESEDA (1)	300	.353	-	.283	=	+.070
RESEDA (2)	300	.597	-	.283	=	+.314
RIVERSIDE (1)	300	.443	-	.383	=	+.060
RIVERSIDE (2)	300	.587	-	.383	=	+.204
TEMPLE CITY (1)	300	.510	-	.297	=	+.213
UPLAND (1)	300	.433	-	.360	=	+.073
UPLAND (2)	300	.577	-	.360	=	+.217

(1) ARB Equation

(2) Modified Equation

It should be noted that only one year of data (1977) was available for the six monitoring sites; hence the verification results are based on a much smaller sample that of oxidant.

4.5 EPISODE PROBABILITIES

Empirical episode probabilities and confidence intervals were computed using the same techniques as described in section 3.5, except that basinwide maxima were used instead of site-specific probabilities. These are given in Table 4.7. (Days falling in the "filtered" category were treated separately.) Thus, if one takes the maximum sulfate value as generated by the equations/modified equations, a probability of an episode ($\geq$ 25 μg/m^3) can be obtained. For example, a maximum sulfate prediction of 24 μg/m^3 has a corresponding 53.8% probability that the observed value will be greater than 25 μg/m^3.

4.6 THIRTY-HOUR PREDICTIONS

The previous modifications to the ARB equations were site specific predictions after today's sulfate value (i.e. the persistence term) is known. To provide additional lead-time for the purpose of assessing the sulfate potential before the persistence term is available, regression equations were generated using basin-maximum sulfate data versus key early-morning meteorological variables.

As in the 24-hour predictions, a "filter" was constructed to eliminate those days in which the likelihood of an episode ($\geq$25 μg/m^3) was quite low. From the unfiltered data, linear, log-linear, and log-log regressions were run. The differences in the results were negligible, so the linear case was retained. To add greater predictability, output from the 24-hour LFM 500 mb progs were used as predictors of several key meteorological variables.

The results are given below:

$$
\begin{aligned}
SO^{=}_{4(\text{basin max})} &= 0.56\ (\text{LAXIDT4}) + 0.22\ (\text{BASNFMY}) \\
&\quad + 1.73\ (\text{LAXITT4}) - 0.74\ (\text{LAX8TM4Y}) \\
&\quad + .0006\ (\text{LAXIBH4}) - 0.19\ (\text{SAN5HT2}_p) - 0.21\ (\text{OAK5HT2}_p) + 20.3
\end{aligned}
$$

Table 4.7 Empirical Episode Probabilities and Confidence Intervals

DERIVED FROM LOG NORMAL ERROR DISTRIBUTIONS

PREDICTED VALUE	RANGE USED	TOTAL IN RANGE	EPISODE PROBABILITY	CONFIDENCE INTERVALS					
				--- 80% ---		--- 90% ---		--- 95% ---	
FILTERED		334	0.6						
15.	1	222	2.2	12.2	21.6	10.8	23.5	9.0	24.9
16.	1	222	5.1	13.0	23.0	11.6	25.0	9.7	26.6
17.	1	222	8.4	13.8	24.4	12.3	26.6	10.3	28.3
18.	1	222	13.8	14.6	25.9	13.0	28.2	10.9	29.9
19.	1	222	20.1	15.4	27.3	13.7	29.7	11.5	31.6
20.	1	222	26.6	16.2	28.8	14.4	31.3	12.1	33.3
21.	1	222	31.6	17.1	30.2	15.2	32.9	12.7	34.9
22.	1	222	38.7	17.9	31.6	15.9	34.4	13.3	36.6
23.	1	222	46.2	18.7	33.1	16.6	36.0	13.9	38.2
24.	1	222	53.8	19.5	34.5	17.3	37.5	14.5	39.9
25.	1	222	65.3	20.3	35.9	18.1	39.1	15.1	41.6
26.	1	222	72.3	21.1	37.4	18.8	40.7	15.7	43.2
27.	1	222	79.5	21.9	38.8	19.5	42.2	16.3	44.9
28.	1	222	81.7	22.7	40.3	20.2	43.8	16.9	46.6
29.	1	222	84.5	23.6	41.7	20.9	45.4	17.5	48.2
30.	1	222	87.6	24.4	43.1	21.7	46.9	18.1	49.9
31.	1	222	91.7	25.2	44.6	22.4	48.5	18.7	51.6

where:

LAXIDT4 = LAX 14Z inversion ΔT (°C)

BASNFM4 = Yesterday's basin-max sulfate ($\mu g/m^3$)

LAXITT4 = LAX 14Z inversion top temperature (°C)

LAX8TM4Y = Yesterday's LAX 850 mb 14Z temperature (°C)

LAXIBH4 = LAX 14Z inversion base height (ft)

$SAN5HT2_p$ = Predicted 24-hour 12Z SAN 500 mb heights
from LFM Prog (10m - 5000)
(i.e. 5<u>86</u>0 m represented by 86)

$OAK5HT2_p$ = Predicted 24-hour 12Z OAK 500 mb height from LFM
Prog (10m - 5000)

The equation is used on days which are not filtered according to:

Predict <25 $\mu g/m^3$ if:

(1) LAX 14Z inversion base is:

(a) surface

(b) ≥5000 ft (includes "no inversion")

(2) LAX 14Z inversion ΔT is ≤2°C

Of the 177 cases (during 1977) used in the equation development, 49 were filtered. For those 49 cases, the average basin-max sulfate value on the following day was 9.8 $\mu g/m^3$ with a maximum value of 22.0 $\mu g/m^3$. Thus no episode conditions occurred on the day following the filtered condition. For the remaining 128 cases, 12 out of 24 episodes were correctly predicted for a 50% probability of detection with only 6 false alarms. The overall verification is as follows:

N	T_c	-	10E	+	T_2	+	P	=	Total
(177)	89		54		31		46		112

The verification results are not as good as the 24-hour site-specific equations, but are appreciably better than one-day persistence. Therefore, this prediction algorithm will provide the ARB with an objective method for preliminary basin-maximum sulfate potential based on early morning data.

CHAPTER 5
SULFUR DIOXIDE PREDICTION TECHNIQUE

5.1 INTRODUCTION

5.1.1 General Methodological Overview

The ambient concentration of SO_2 is affected by both SO_2 emissions and meteorologically-induced dilution and dispersion. Previous SO_2 models concentrated on available SO_2 emissions data (Chapter 2), simulating occurrences of high SO_2 levels for which emissions data were unavailable. The goal of this project was to generate real-time meteorologically-based SO_2 prediction algorithms capable of predicting SO_2 better than either persistence or climatology alone, but not requiring detailed real-time emissions data, which are generally not available.

The meteorological dispersion of an emission from an identified source can be expressed by the vertical and horizontal dispersion of the emissions. In the SCAB, vertical mixing is often limited by a persistent marine inversion. In general, the base of the inversion acts as a lid, defining the height of the layer in which a pollutant mass (an SO_2 plume) can be mixed. The strength of this lid is expressed by the intensity of the inversion (i.e., the change in temperature between the base and top of the inversion). The concentration of SO_2 downwind is inversely related to the amount of vertical mixing.

Lateral dispersion is basically a function of the wind-driven dilution and gradient forcing - both thermal and pressure. The wind direction and speed are key variables that define potential areas affected by emissions from a given source or group of sources. Pressure and thermal gradients act as forcing mechanisms for the wind field and indicate possible stagnation situations.

All parameters in the data base (described in Chapter 2) were used as possible predictors. Relationships between upwind indicators and selected local variables were examined to establish any clues in the changes of small-scale features (e.g. the inversion base height), and synoptic scale features (e.g. a Santa Ana wind condition), all having potential effects on tomorrow's SO_2. By combining the contributions of both the long range and local met variables, a compilation of the most important meteorological predictors of SO_2 was created.

This section of the report details the methods to identify the best and most probable meteorological predictors of SO_2 and to establish a real-time set of forecast algorithms. An evaluation of the basinwide distribution of SO_2 concentrations together with a verified prediction scheme will also be presented. To evaluate the contribution of emissions data for the forecast algorithm a separate case study was performed to determine the influence of the Haynes and Los Alamitos power plant emissions on the ambient SO_2 levels at Los Alamitos.

5.1.2 Los Angeles SO_2 Distribution

In the Los Angeles basin, SO_2 daily maximum one hour average concentrations typically range from near 0 to approximately 20 pphm. Values above 20 pphm are rare and values above the one hour maximum standard of 50 pphm have not been recorded in several years. Therefore, the objective of these SO_2 prediction algorithms is to detect days whose 1-hour average max concentrations exceed 10 pphm. This threshold was chosen because a combined episode condition exists when both SO_2 and OX average concentrations exceed 10 pphm in the same one hour period.

Ambient levels of SO_2 reflect transport from source (SO_2 emission) areas to receptor areas. The potential for high SO_2 levels varies within the basin, as shown by the spatial display of mean one-hour max SO_2 isopleths (Figure 5.1). The largest concentrations occur in and downwind of the oil refineries and power plants dotting the coast south of LAX. The second area of high SO_2 is Fontana, where emissions from steel industries and power plants are high. Also, since Fontana is a high-ozone site, the likelihood of combined oxidant/sulfur dioxide episodes is greater there than for the coastal sites. By relating the regional values of SO_2 to the values observed at Fontana and Lennox (for the Coast) and a Basin-Max, a basic network of SO_2 prediction was established.

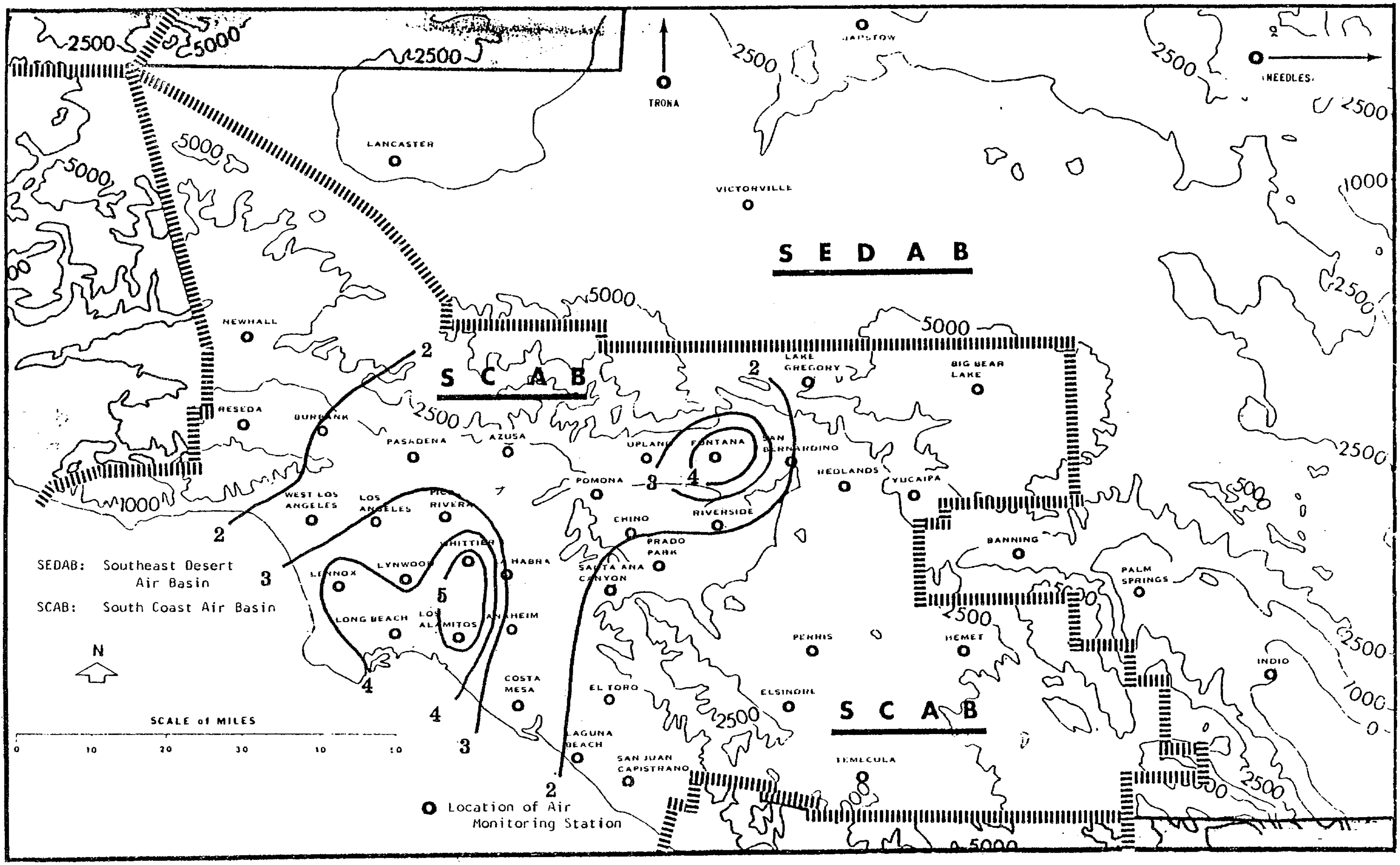


Figure 5.1 Mean One Hour Max SO_2 Isopleths for the L.A. Basin (Isopleth in pphm)

The main emphasis of the SO_2 prediction development was to predict concentrations at Fontana and Lennox, each of which represents one of the two problem areas. Other stations measuring SO_2 were linked to Fontana, Lennox or the Basin-Max by means of regression analysis. As will be discussed later, some stations with persistently low values were better predicted by climatology than any other method.

5.1.3 Basinwide Distribution

Initially, to determine those sites with significant SO_2 values, an examination of the concentration distribution for each station was performed. Stations displaying lower values of observed SO_2 were examined for the use of either persistence or climatology as a prediction method. Other stations showing lower values of SO_2 (under 10 PPHM with occasional higher values) were examined to see whether a regression equation to the predicted values of Lennox and/or Fontana was necessary for good prediction accuracy. Results showed that it was necessary to relate SO_2 at Long Beach, Lynwood, La Habra, Los Alamitos, Whittier, Costa Mesa and Anaheim to the predicted values of SO_2 at Lennox and Fontana.

A review of the individual daily max SO_2 values and the monthly average of the max concentrations indicated that several stations never experienced values of SO_2 equal to or greater than 10 pphm for the three years of the study period. Since the evaluations of the prediction accuracy were split into categories where 10 pphm acted as the initial threshold for Reseda, Pomona, Pasadena-Walnut and Riverside-Rubidoux, the use of climatology as a predictor achieved 100% categorical prediction accuracy by predicting zero exceedance of 10 pphm threshold. Monthly means of the daily max concentrations were often between 1 and 3 pphm while few values exceeded 5 - 6 pphm. A majority of the SO_2 values for each month fell within ± 2 pphm of the mean. An estimated percentage of absolute error of prediction was also low. In general for these stations either climatology or persistence would act as the best predictor. Improvement over these algorithms would be slight, if any.

Stations exhibiting occasional SO_2 concentrations exceeding the 10 PPHM threshold included: Newhall, West L.A., DOLA, Azusa, and Burbank. These stations experienced daily max SO_2 concentrations above the threshold less than 6 times over the four year period 1974-1977. Climatology achieved nearly perfect prediction accuracy with a majority of predictions falling within ± 2 pphm of observed. Persistence was less effective because of the rarity in the occurrence of back-to-back days having SO_2 1 hr max concentrations $\geq$ 10 pphm.

Stations not using climatology as a prediction method were related to either Lennox or Fontana by means of regression. Table 5.1 is a complete list of stations, their corresponding prediction methodology, and set of regression equations relating each station to a combination of forecast SO_2 values including Fontana, Lennox and the Basin-Max.

5.1.4 Prediction Algorithms

Prediction algorithms for Fontana, Lennox and the Basin-Max, consisted of same day, same day (8-11 A.M.), 24 hour and 30 hour forecasts of SO_2. Two main statistical methods were used to determine valid forecast algorithms: stepwise multiple regression and computer aided pattern recognition (AID). Empirical techniques utilizing individual expertise including point classification systems were also implemented. In each case, the 1974-76 SO_2 data set was tested against independent meteorological variables and persistence for two separate seasons - summer (May - Oct), and winter (Nov - April). The analysis resulted in a series of algorithms having distinct prediction capabilities.

In Fontana, the daily 1-hr average max concentrations of SO_2 often coincide with the 1-hr max between the hours of 8 A.M. and 11 A.M. The correlation coefficient between the two max hourly values is approximately 0.94. As a result, same day prediction algorithms were focused upon prediction for the (8-11) A.M. period.

The significance of forecasting SO_2 for the (8-11) A.M. period is to determine the potential of a violation of the combined SO_2 - oxidant episode criteria. Oxidant values are building up during this period while SO_2 is

Table 5.1 SO_2 Prediction Equations for SCAB Sites as Functions of Key Predictor Sites

Station	Prediction Equation	Number Cases: N	R	R^2	Standard Error
DOLA	DOLA = 0.18 LENX + 0.12 BASN + 1.41	921	0.57	0.32	1.28
LA HABRA	LAHB = 0.27 BASN + 0.18 LENX - 0.01	921	0.59	0.35	1.94
BURBANK	BURK = 0.13 LENX + 0.09 BASN + 1.04	921	0.44	0.19	1.29
WEST L.A.	WEST = 0.15 LENX + 0.05 BASN + 1.17	921	0.46	0.21	1.13
LONG BEACH	LONB = 0.49 BASN + 0.07 LENX + 0.29	921	0.64	0.41	2.41
NEWHALL	NEWH = 0.05 LENX + 0.03 BASN + 1.42	1057	0.20	0.04	1.12
RESEDA	RESD = 0.06 LENX + 0.03 BASN + 0.98	1057	0.35	0.12	0.79
PASADENA	PASD = 0.10 LENX + 0.06 BASN + 1.48	1057	0.50	0.25	0.85
AZUSA	AZUS = 0.14 LENX + 0.07 BASN + 1.20	1057	0.50	0.25	1.07
POMONA	POMA = 0.12 BASN + 0.13 LENX + 0.92	1057	0.56	0.31	1.13
WHITTIER	WHTR = 0.43 BASN + 0.22 LENX + 0.78	720	0.65	0.43	2.43
ANAHEIM	ANAH = 0.22 BASN + 0.03 LENX + 0.56	720	0.46	0.21	1.77
COSTA MESA	COST = 0.19 BASN + 0.14 LENX + 0.60	720	0.48	0.23	1.90
LYNWOOD	LYND = 0.30 LENX + 0.15 BASN + 1.21	720	0.62	0.39	1.62
LOS ALAMITOS	LSAL = 0.94 BASN - 0.25 FONT - 0.19 LENX - 0.42	464	0.72	0.51	2.83
SAN BERNARDINO	SNBD = 0.15 FONT + 1.51	527	0.32	0.10	1.62
RIVERSIDE/RUBIDOUX	RIVR = 0.20 FONT + 0.08 LENX + 0.70	475	0.56	0.31	1.22
SANTA ANA CANYON	SACN = 0.14 BASN - 0.07 LENX - 0.30	161	0.51	0.27	0.94

generally peaking. Although oxidant will usually peak during the early afternoon, it often exceeds 10 pphm during this period, causing the possibility of a violation.

5.2 SAME DAY FORECASTS

Same day forecasts were produced for the two key stations (Lennox and Fontana, 8-11 A.M.) and for the Basin-Max. Additional same day forecast algorithms were developed for Whittier and Los Alamitos. For Fontana and Lennox, AID-created decision-trees produced the best forecast algorithms. Although the decision tree analysis gave a discrete pollution prediction as opposed to the continuous prediction capabilities of regression analysis, it proved to be an adequate method of forecasting SO_2.

An empirically derived classification system was determined to predict the Basin-Max while regression analysis was used as an alternate prediction system for Lennox. The prediction verification scores will be presented in Section 5.2.3.

Key variables used by the different algorithms include persistence and predicted values of Upland oxidant for the Fontana algorithms (both seasons). Lennox decision trees depended heavily upon persistence and LAX inversion variables. The Basin-Max algorithm was based solely upon the values of the 850 mb and 950 mb temperatures and persistence.

5.2.1 Final Forecast Algorithms

Presented in this section are the final forecast algorithms developed for Lennox, Fontana and the Basin Max.

(1) Lennox

(See decision trees Figure 5.2 (summer) and Figure 5.3 (winter).)

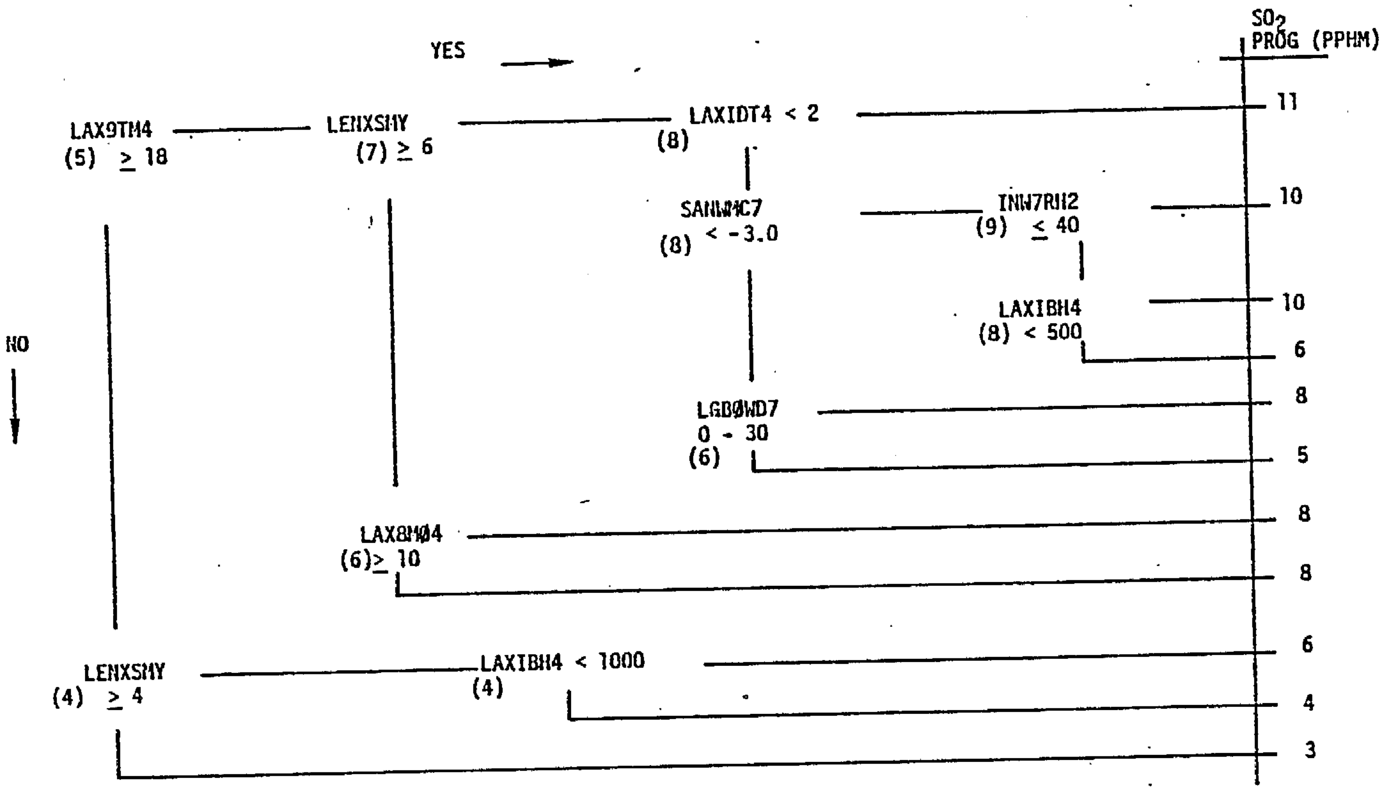
(2) Fontana

(See decision trees Figure 5.4 (summer) and Figures 5.5 and 5.6 (winter).)

(3) Basin-Max

All Year

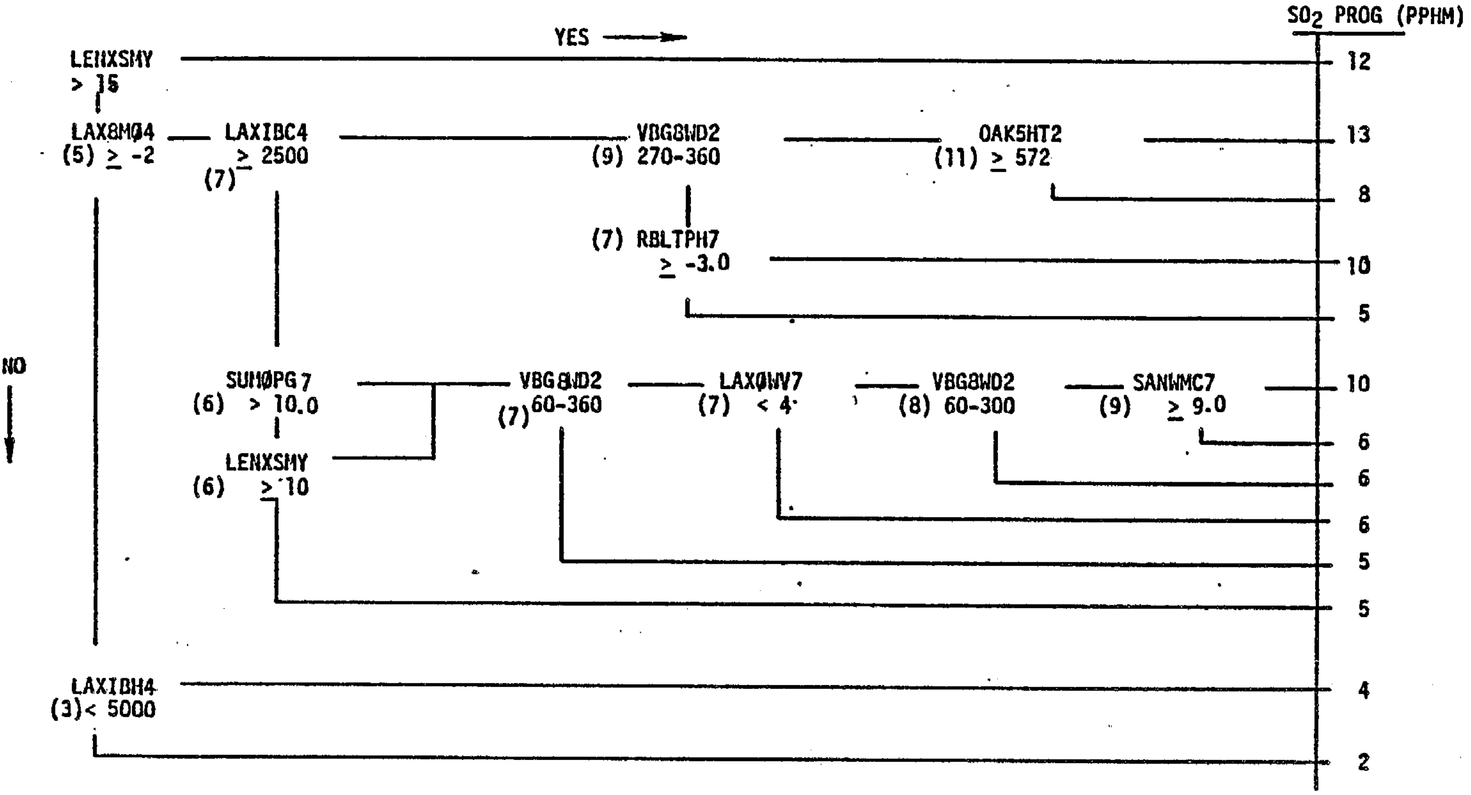
BASNSMO = ,5 (.5 (LAX8TM4 + (LAX8DIF + LAX9DIF)) + BASNSMY)



(NOTE: Number encircled, ie (7), indicates predicted value if data are not available for further splits.)

Where:	LAX9TM4	- 7 a.m. 950 mb temp at LAX (°C)
	LENXSMY	- Yesterday's SO_2 1-hr max at Lennox (PPHM)
	LAXIDT4	- 7 a.m. inversion top temp - base temp at LAX (°C)
	LAX8MØ4	- 7 a.m. 850 mb temp - surface temp at LAX (°C)
	SANWØC7	- 7 a.m. pressure gradient between SAN - WMC (mb)
	INW7RH2	- 700 mb relative humidity at INW (%)
	LAXIBH4	- 7 a.m. LAX inversion base height (Ft)
	LGBØWD7	- 7 a.m. surface wind direction at LGB

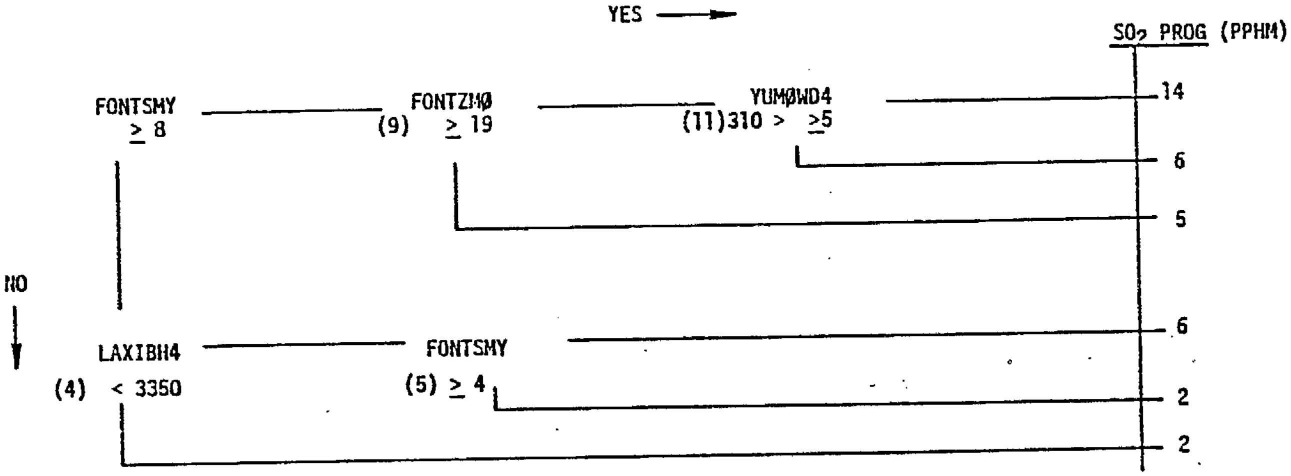
Figure 5.2 Same Day Summer SO_2 Predictions for Lennox, $R^2 = 0.51$



(NOTE: Number encircled, ie (7), indicates predicted value if data are not available for further splits.)

Where:
LENXSMY - Yesterday's 1-hr SO_2 max at Lennox (PPHM)
LAX8MØ4 - 7 a.m. 850 mb temp-surface temp at LAX (°C)
LAXIBC4 - 7 a.m. inversion thickness at LAX (Ft)
LAXIBH4 - 7 a.m. LAX inversion base height (Ft)
VBG8WD2 - 850 mb wind direction at VBG (12 Z)
OAK5HT2 - 500 mb height at OAK - 12 Z (10 M)
LAXØWV7 - 7 a.m. surface wind velocity at LAX (MPH)
SANWMC7 - 7 a.m. pressure gradient between SAN-WMC (mb)
RBLTPH7 - 7 a.m. pressure gradient between RBL-TPH (mb)
SUMØPG7 - Σ 7 a.m. pressure gradients: LAX-DAG, SAN-LAS- SDB-VCV (mb)

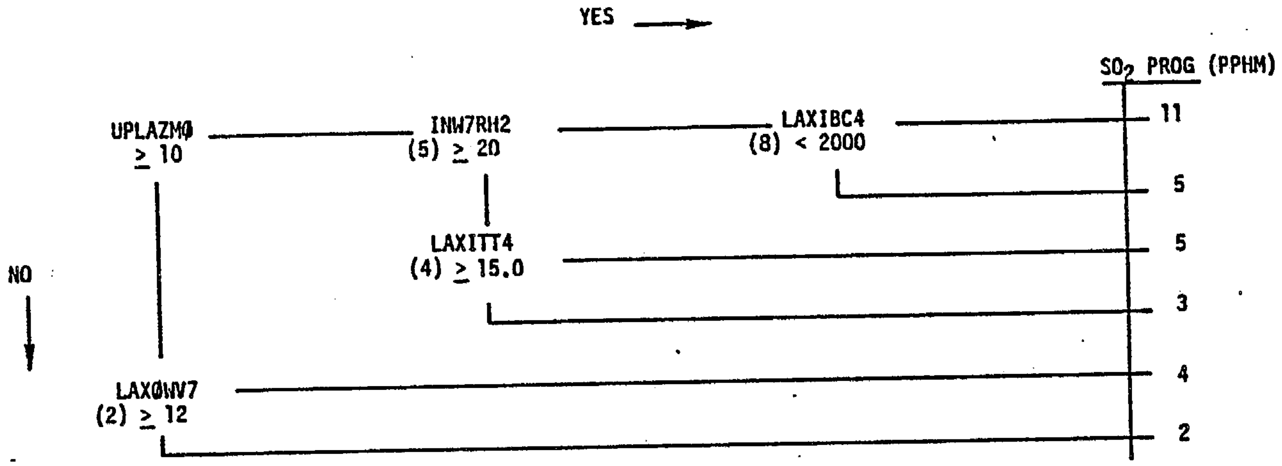
Figure 5.3 Same Day Winter SO_2 Predictions for Lennox, $R^2 = 0.52$



(NOTE: Number encircled, ie (9), indicates predicted value if data are not available for further splits.)

Where: FONTSMY - Yesterday's 1-hr SO_2 max at FONT (PPHM)
FONTZMØ - Today's (predicted - observed) ozone 1-hr max at FONT (PPHM)
YUMØWD4 - 4 a.m. surface wind direction at YUMA
LAXIBH4 - 7 a.m. LAX inversion base height (Ft)

Figure 5.4 Same Day Summer SO_2 Predictions (8-11 a.m.) for Fontana, $R^2 = 0.73$



NOTE: Number encircled, ie. (5), indicates predicted value if data are not available for further splits.)

Where: UPLAZMØ - Today's predicted 1-hr max ozone at UPLA (PPHM)

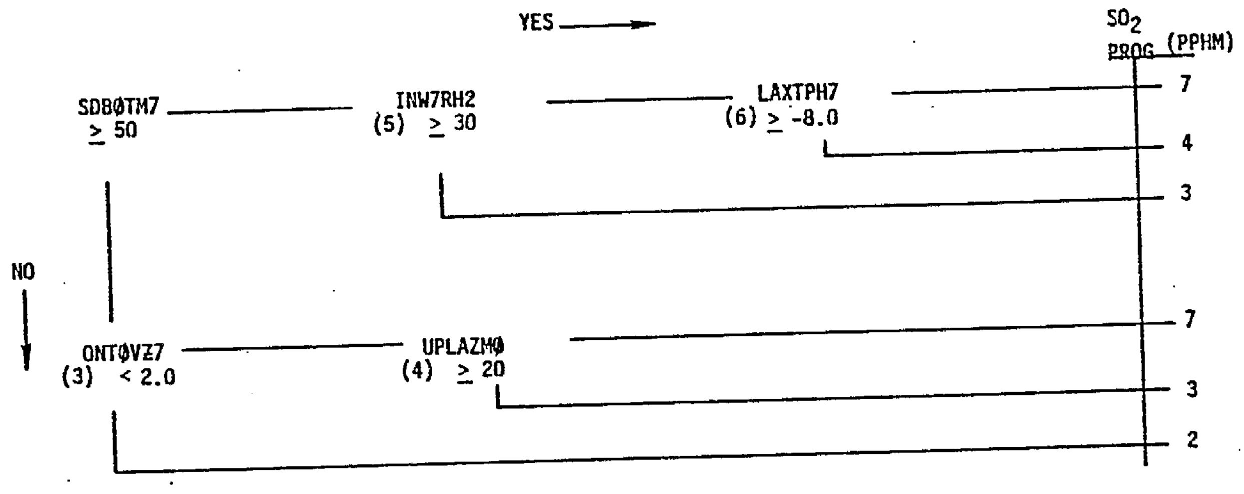
INW7RH2 - 700 mb relative humidity at INW (%)

LAXIBC4 - 7 a.m. LAX inversion thickness top height - base height (Ft)

LAXITT4 - 7 a.m. LAX inversion top temp

LAXØWV7 - 7 a.m. surface wind velocity at LAX (mph)

Figure 5.5 Same Day Winter SO_2 Predictions (8-11 a.m.) for Fontana with LAXIBH4 ≠ 110, R^2 = 0.69



(NOTE: The number encircled, ie (5), indicates predicted value if data are not available for further splits.)

Where: SDBØTM7 - 7 a.m. surface temp at SDB (°F)
ONTØVZ7 - 7 a.m. surface visibility at ONT (miles)
INW7RHZ - 700 mb relative humidity at INW, 12Z (%)
UPLAZMØ - Today's predicted 1-hr max ozone at UPLA (PPHM)
LAXTPH7 - 7 a.m. pressure gradient between LAX - TPH (mb)

Figure 5.6 Same Day Winter SO_2 Predictions (8-11 a.m.) for Fontana with LAXIBH4 = 110, R^2 = 0.79

where LAX8DIF = LAX8TM4 - LAX8TMY
LAX9DIF = LAX9TM4 - LAX9TMY
LAX9TMY = 7 A.M. 950 mb Temp at LAX (°C)
LAX8TMY = 7 A.M. 850 mb Temp at LAX (°C)
BASNSMY = Yesterday's 1-hour max SO_2 (pphm)

$$\text{NOTE: } (\text{LAX9DIF} + \text{LAX8DIF}) \begin{array}{l} \geq 10.0 = 10.0 \\ \leq -10.0 = -10.0 \end{array}$$

(4) Alternate Lennox Same Day Equations

Summer Base Ht = 110 N=38 R=0.65 SE=2.63

$$\begin{aligned} \text{LENXSMØ} = & - 0.05\ \text{LAXIDT4} + 0.30\ \text{LENXSMY} - 0.17\ \text{SUMØPG7} \\ & + 0.16\ \text{LAXØTM4} + 0.28\ \text{LAX8MØ4} + 0.30\ \left(\frac{\text{LAXØWV4} + \text{LAXØWV7}}{2}\right) \\ & + 1.81 \end{aligned}$$

Summer Base Ht ≠ 110 N=440 R=0.60 SE=2.09

$$\text{LENXSMØ} = 0.38\ \text{LENXSMY} - 0.00045\ \text{LAXIBH4} - 0.07\ \text{SUMØPG7} + 4.63$$

Winter Base Ht = 110 N=229 R=0.52 SE=2.87

$$\begin{aligned} \text{LENXSMØ} = & \ 0.36\ \text{LAX8MØ4} - 0.18\ \text{SUMØPG7} + 0.20\ \text{LENXSMY} \\ & - 0.00022\ \text{LAXIBC3} + 5.83 \end{aligned}$$

Winter Base Ht ≠ 110 N=224 R=0.65 SE=2.04

$$\begin{aligned} \text{LENXSMØ} = & \ 0.15\ \text{LAX8MØ4} - 0.18\ \left(\frac{\text{LAXØWV4} + \text{LAXØWV7}}{2}\right) \\ & - 0.00024\ \text{LAXIBH4} + 0.11\ \text{LENXSMY} - 0.049\ \text{SUMØPG7} \\ & + 6.64 \end{aligned}$$

Where: LAXIDT4 - 7 a.m. LAX inversion top temp - base temp (°C)
LENXSMY - Yesterday's 1-hour max SO_2 at Lennox (PPHM)
SUMØPG7 - Σ15Z pressure gradients: LAX-DAG, SDB-VCV, SAN-LAS (mb)
LAX8MØ4 - 7 a.m. 850 mb temp - surface temp at LAX (°C)
LAXØTM4 - 7 a.m. surface temp at LAX (°C)
LAXØWV4 - 7 a.m. surface wind speed at LAX (MPH)
LAXØWV7 - 7 a.m. surface wind speed at LAX (MPH)
LAXIBH4 - 7 a.m. LAX inversion base height (Ft)
LAXIBC3 - 7 a.m. 24 hour LAX inversion base change (Ft)

5.2.2 Development of the Forecast Algorithm

This section is a chronological review of the development of the final forecast algorithms. Because of the site specific nature of SO_2, differing procedures were used to determine potential prediction algorithms at the various sites.

Lennox

Initially for Lennox, high SO_2 days (1 hr maxima $\geq$ 10 PPHM) were separated from the dependent data set distribution. Using the SO_2 values of this data subset, a manual examination of SO_2 data versus meteorological parameters was conducted to gain a "feel" for the situation. By fitting SO_2 values to various meteorological parameters and persistence a limited prediction algorithm was made. Since the effort was geared to high SO_2 values, it tended to overpredict when applied to the full data set. Although several met variables showed definite promise as possible predictors, the number of potential predictors was too large for a continued manual fitting.

Regression analysis was then used to predict SO_2 for same day algorithms. SO_2 variables were related by stepwise multiple regression to locally selected parameters based on known success as oxidant prediction. (The oxidant prediction variables represent a statement of the meteorological potential for high pollution on a given day.) Then, local wind velocity variables and specific pressure gradients were added, step by step to improve the developing algorithms. The resulting equations were broken into two sets of categories according to the use of persistence and the presence of a surface inversion for each season.

Among the equations produced by the regression analysis, the cases which included persistence were the best predictors. The resulting four equations represented winter and summer predictions for days when the LAX inversion base height was either equal to 110 ft (surface inversion) or greater. Using this criterion we were able to stratify the better potential predictors according to the meteorological situation (i.e., with a surface inversion present, inversion strength variables acted as the best predictors).

With a series of algorithms being produced it was necessary to quickly evaluate the prediction accuracy of each model. Two methods of determining accuracy were to examine the amount of variance explained by the model and to compare the trends of the predicted SO_2 with those of the observed values.

This method used on the Lennox equations illustrated that for summer (LAXIBH4 = 110 ft.) the inclusion of persistence increased the amount of variance explained from $R^2 = 0.37$ to $R^2 = 0.42$. This statistical measure of increased prediction capacity was substantiated by visible improvements in the trends of the predicted SO_2 compared to the observed values.

To enhance the capabilities of the regression analyses a linear matrix that related the SO_2 at Lennox to every variable in the data base was created to determine an optimal set of potential predictors. These variables were combined with selected variables from the previous regression attempts. For the ensuing analysis, a greater emphasis was also placed upon the model's prediction sensitivity, with special attention focused upon the ability to forecast SO_2 concentrations equal to or above 10 PPHM. Additional attention was given to increase the model's ability to catch significant changes in the SO_2 distribution.

Persistence did enhance the prediction capabilities of the previous equations, however, it tended to cause the equation to underpredict higher SO_2 values, particularly on significant change days. To increase the sensitivity of the algorithm, a new series of regression equations was developed without the input of persistence. The resulting algorithms predicted with about the same accuracy as the previous equation set.

The results of this analysis tended to confirm some of the pitfalls encountered with the use of the regression analysis. The regression equation is essentially a best fit of a series of linear variables describing a predictand. The Lennox regression equations accurately predicted SO_2 values near the mean of the distribution but generally missed higher concentrations. Another poor feature of the use of linear regression was elimination of many possible key variables that were related to SO_2 non-linearly. As a result several variables including the wind directions were eliminated from consideration for the equations.

The next step, therefore, was to use AID to develop decision-tree algorithms. By scanning the previous attempts at prediction algorithms as well as the correlation tables, a set of optimal predictors was established for Lennox. This set included local and synoptic wind conditions in discrete intervals as potential predictors. Persistence was also included in the development of the AID algorithms. The most important variables were the difference between the 850 temp and the surface temp at LAX for the winter and the 950 mb temp at LAX for the summer. In the winter algorithm the 850 mb Vandenberg 12Z and the 7 A.M. surface wind directions were also significant, while the surface 7 A.M. wind speed at LGB contributed to the summer AID tree.

Fontana

Same-day prediction algorithms for Fontana were determined using a similar method. A preliminary regression analysis, based solely upon meteorological variables, was conducted for both SO_2 and oxidant to estimate the predictability of SO_2 relative to oxidant. Results showed that for the three years, using the same meteorological variables, more variance was explained for oxidant ($R^2 = 0.59$, for 406 cases) than for SO_2 ($R^2 = 0.48$ for 212 cases).

The SO_2 equations determined from this analysis were not effective as prediction algorithms. Equations developed were derived using a limited set of valid data and did not apply when tested against the full data set. The equations also suffered from an inability to predict values of SO_2 greater than 10 PPHM.

Additional regression attempts were made for Fontana with specific attention paid to highly correlating linear variables. Using the set of best linear meteorological predictors as well as selected long range predictors and persistence, various algorithms were developed. In general, the amount of variance described by this series of equations was reduced from that of the initial equation (as a result of a larger number of test cases to be fit). One of the initial regression equations developed related SO_2 at Fontana to long range variables including UCC5HT2, MFRWMC7, RBLTPH7, and local variables such as SUMØPG7 and SDBØWV7. (This regression equation achieved an $R^2 = 0.36$, 202 summer cases.) Models were related to both the daily max SO_2 and the (8-11) A.M. average. For all cases, the resulting correlations were poor, with particularly inaccurate SO_2 predictability in the high range.

Following the failure of the regression analysis to produce a working algorithm, decision tree analysis was initiated. Initial runs tried to relate Fontana (8-11) A.M. daily max hourly SO_2 (with restriction of OX $\geq$ .10 for that day) to different meteorological variables. For the same-day trees, a majority of the predictor variables were LAX inversion parameters. The key variables in the (8-11) A.M. forecast were LAX8TM4, SDBØTM7 and ONTØVZ7. Only one long range synoptic variable appeared in the trees, Oakland's 500 mb height change (12Z).

The initial decision tree was unable to predict values of $SO_2 \geq$ 10 pphm. However, the amount of variance explained for the (8-11) A.M. forecast was relatively low ($R^2 = 0.31$).

To continue the decision tree approach, a complete survey of possible predictors was performed to determine an optimal set. Decision trees were constructed for the two main forecast periods: 8-11 A.M. summer, 8-11 A.M. winter with the LAX inversion greater than 110 ft., and 8-11 A.M. winter with a surface inversion. The resulting algorithms proved to be adequate forecast tools. Minor revisions were made in each algorithm to simplify the output without losing any resolution in the actual forecast.

The final algorithms explained a high percentage of variance in the Fontana SO_2 distribution. The summer algorithm had an $R^2 = .73$, while the winter algorithms explained 75 percent of the variance.

Basin-Max

To determine the same day Basin-Max SO_2 two major methods were attempted: regression and an empirical meteorological potential system derived through interactive analysis.

The regression analysis produced separate equations for the summer and winter, each predicting better than persistence. However, the model predicted consistently low for most values of the Basin-Max. Improvements over persistence were almost insignificant. The model did predict mid-range values of the Basin Max accurately, but it was unable to predict the magnitude of high SO_2 concentrations.

As a part of the development of an oxidant model for Riverside, an algorithm was derived expressing the OX as a function of the 850 mb temperature, the 950 mb temperature, and the differences of those variables over the past 24 hours. The resulting model produced an accurate forecast not only for oxidant but for the meteorological potential of pollutants for that day. Since the Basin-Max SO_2 concentration is more a function of a regionwide potential than a site-specific local phenomenon, we attempted to fit the oxidant - meteorology algorithm to the Basin-Max distribution.

Basin-Max oxidant values for average days are generally twice the Basin Max SO_2 values for that same day (approximately 21 PPHM compared to 9 PPHM). By taking one half of the oxidant forecast based solely on the meteorological potential described by the 850 mb and 950 mb temperatures, and averaging that value with yesterday's Basin-Max, a simple yet highly accurate model was produced.

The model itself was generally conservative in that it predicted somewhat high for a majority of the cases. The forecast SO_2 values tracked the observed values surprisingly well. One benefit of this conservative model is that in the basinwide distribution most stations' SO_2 values are determined by a combination of the Basin-Max and either Lennox or Fontana. With a conservative forecast possible local peaks will be more efficiently detected.

5.2.3 Verification Analysis - Same Day Forecasts

5.2.3.1 Verification Against Dependent Data

All algorithms were verified against the dependent data set to compare their forecast abilities against persistence. Table 5.2 presents summer and

Table 5.2 Overall Same-Day SO_2 Prediction Rating

N	METHOD	T_c	- 10E	+ T_2	+ C	+ P	= R	R/PERFECT R
	PERFECT	100	0	100	100	25	325	
	DEPENDENT DATA SET: 1974-1976							
	MAY-OCT							
329	Fontana-persistence	84	26	61	0	25	144	.443
543	Lennox-persistence	91	19	71	0	25	168	.517
557	Basin-Max-persistence	73	29	55	0	25	125	.385
266	Fontana (8-11 a.m. algorithm)	96	19	67	40	25	209	.643
434	Lennox (algorithm)	93	15	83	37	25	223	.686
556	Basin-Max (algorithm)	75	24	67	23	25	166	.511
549	Lennox climatology	93	23	70	40	25	208	.640
333	Fontana climatology	86	30	50	37	25	168	.517
	NOV-APR							
296	Fontana-persistence	93	23	67	0	25	162	.498
527	Lennox-persistence	86	26	61	0	25	146	.449
541	Basin-Max-persistence	71	27	56	0	25	125	.385
216	Fontana (8-11, both algorithms)	97	12	88	74	25	272	.837
397	Lennox (algorithm)	90	17	73	35	25	206	.634
456	Basin-Max (algorithm)	71	27	59	36	25	164	.505
537	Lennox climatology	91	23	64	40	25	197	.606
304	Fontana climatology	96	22	68	48	25	215	.602
	INDEPENDENT DATA SET: 1977							
	MAY-OCT							
141	Fontana-persistence	86	32	50	0	25	129	.397
182	Lennox-persistence	97	18	81	0	25	185	.569
184	Basin-Max-persistence	63	34	52	0	25	106	.326
141	Fontana (8-11 algorithm)	83	33	63	26	25	164	.505
184	Lennox (algorithm)	97	15	84	30	25	221	.680
183	Basin-Max (algorithm)	73	27	59	15	25	145	.446
	NOV-APR							
179	Lennox-persistence	89	17	73	0	25	170	.523
179	Lennox (algorithm)	94	18	79	42	25	222	.683
180	Basin-Max-persistence	82	26	66	0	25	147	.452
178	Basin-Max (algorithm)	75	24	65	41	25	182	.560

LEGEND

* = Best Method
N = Number of Predictions
T_C= Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes
Correct ±2 PPHM (%)

R = Rating
P = Score (Climatological
Constant)

winter verification scores. Model evaluation was performed using the basic scoring system described in Chapter 2, with one alteration: significant change days were defined to be days where SO_2 increased or decreased 5 pphm from the preceding day. A perfect score for the model was 325 points. (For all models P=25 (the equivalent of climatology) because concentrations of 1-hr SO_2 never exceeded 50 pphm at any station over the modeling period.)

It can be seen that for the three algorithms, improvement was made over persistence and climatology at each site. The Fontana 8-11 A.M. winter algorithm achieved a score of 272, which is 83.7% of perfect. This represents the best single algorithm developed for any of the pollutants in this study.

5.2.3.2 Verification on Independent 1977 Data

All prediction algorithms except Fontana-winter were evaluated against 1977 SO_2 and meteorological data. The algorithms were evaluated using all available data. Results of the independent analysis are also shown in Table 5.2,

Similar to the dependent data set, each of the new algorithms scored substantially better than persistence, indicating that these methods can be used as effective prediction tools in subsequent years.

5.3 24-HOUR DAY IN ADVANCE PREDICTIONS

Twenty-four hour forecast algorithms for Lennox, Fontana and the Basin-Max were restricted to computer derived decision trees and regression analyses. The 24-hour forecast algorithms provide an update of the initial 30-hour forecast. The update forecasts are able to use meteorological and pollution data not available in the morning.

For Lennox and Fontana (both seasons) decision trees produced the best predictive algorithms. Key variables for the Lennox winter algorithm include, OAK5HT2 and persistence, and for the summer algorithm persistence and the LAXIBH4 were the most important variables. The 24-hour forecast decision trees designed for Fontana used persistence and predicted oxidant values for Upland as the key variables for both algorithms.

To produce a 24-hour algorithm for the Basin Max a series of regression equations were developed. Key variables in the winter equation are OAK5HT2 and SUMØPG7. Variables weighing heavily in summer equation are LAXIBH4 and LAX9TM4.

5.3.1 Final Forecast Algorithms

Presented in this section are the final 24-hour forecast algorithms:

(1) Fontana

(see decision trees Figures 5.7 and 5.8.)

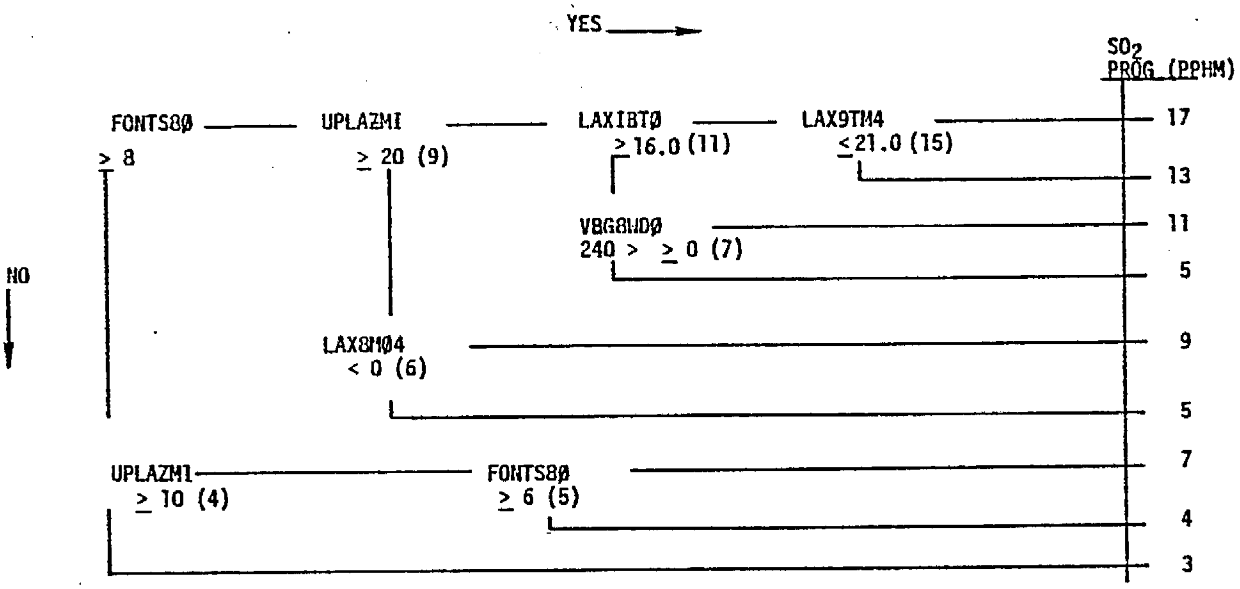
(2) Lennox

(see decision trees Figures 5.9 and 5.10.)

(3) Basin-Max

Summer: $$\text{BASNSM1} = -0.00048\ \text{LAXIBH4} + 0.078\ \text{LAX9TM4} - 4.6\ \text{LAXDAG3} + 0.13\ \text{LAX8TMØ} + 1.4\ \text{SUMØPG7} + 5.6$$

Winter: $$\text{BASNSM1} = 0.07\ \text{OAK5HT2} + 0.14\ \text{LAX8TMØ} - 0.068\ \text{SUMØPG7} - 36.9$$



(NOTE: The number encircled, ie. (9), indicates predicted value if data are not available for further splits.)

Where: FONTS8Ø - Today's (8-11 a.m.) max SO_2 at FONT (PPHM)

UPLAZMI - Tomorrow's predicted 1-hr max ozone at UPLA (PPHM)

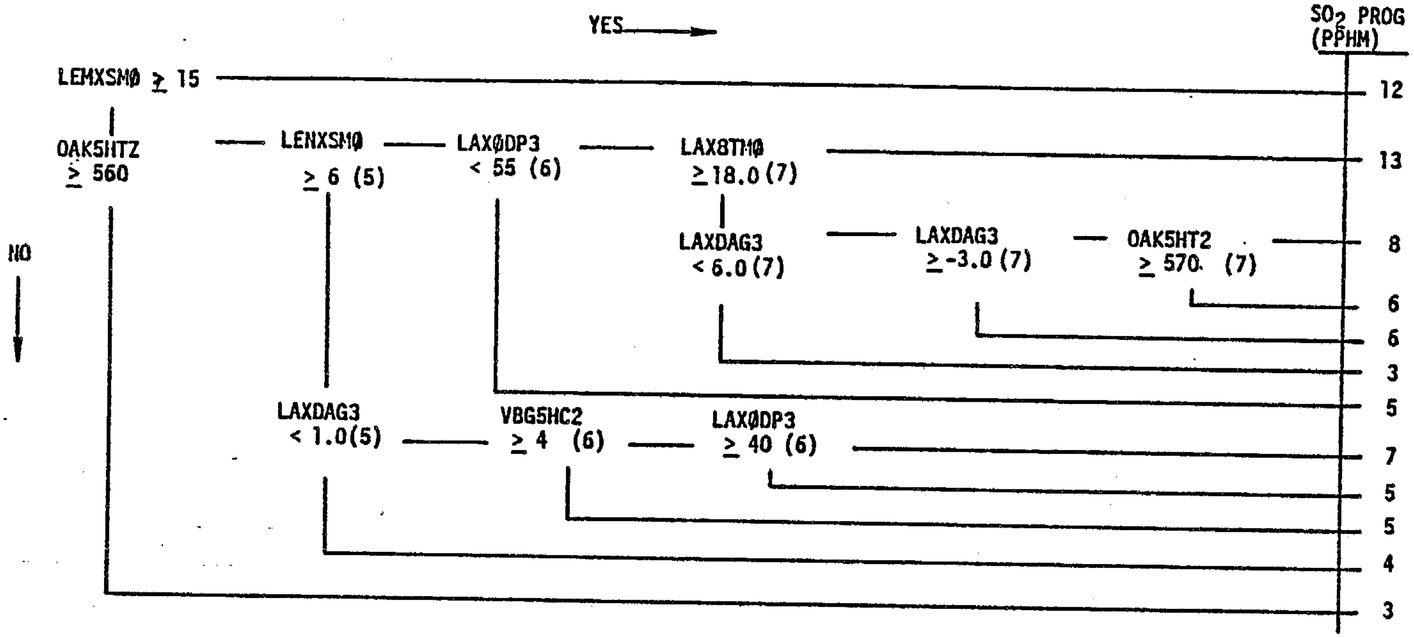
LAXIBTØ - 1 p.m. LAX inversion base temp (°C)

LAX9TM4 - 7 a.m. 950 mb temp at LAX (°C)

VBG8WDØ - 850 mb wind direction at VBG 00Z

LAX8MØ4 - 7 a.m. 850 mb temp - surface temp at LAX (°C)

Figure 5.7 24-Hour Summer SO_2 Predictions for Fontana, $R^2 = 0.69$



(NOTE: Number encircled, ie. (5), indicates predicted value if data are not available for further splits.)

Where: LENXSMØ - Today 1-hr max SO_2 at Lennox (PPHM)

OAK5HT2 - 500 mb heights at OAK 12 Z (10 m)

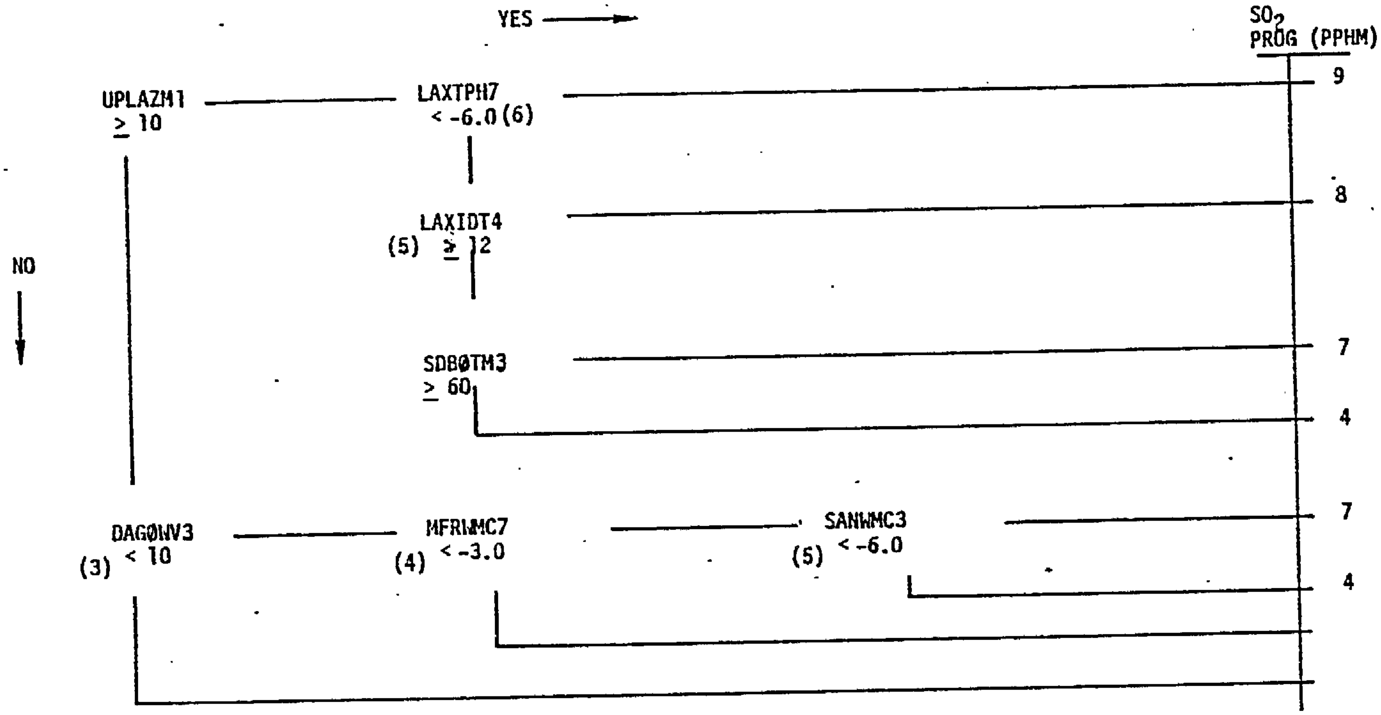
LAXØDP3 - 1 p.m. surface dew point at LAX (°F)

LAX8TMØ - 1 p.m. 850 mb temp at LAX (°C)

LAXDAG3 - 1 p.m. pressure gradient between LAX-DAG (mb)

VBG5HC2 - 24-hour 500 mb height change at VBG 12 Z, (10 m)

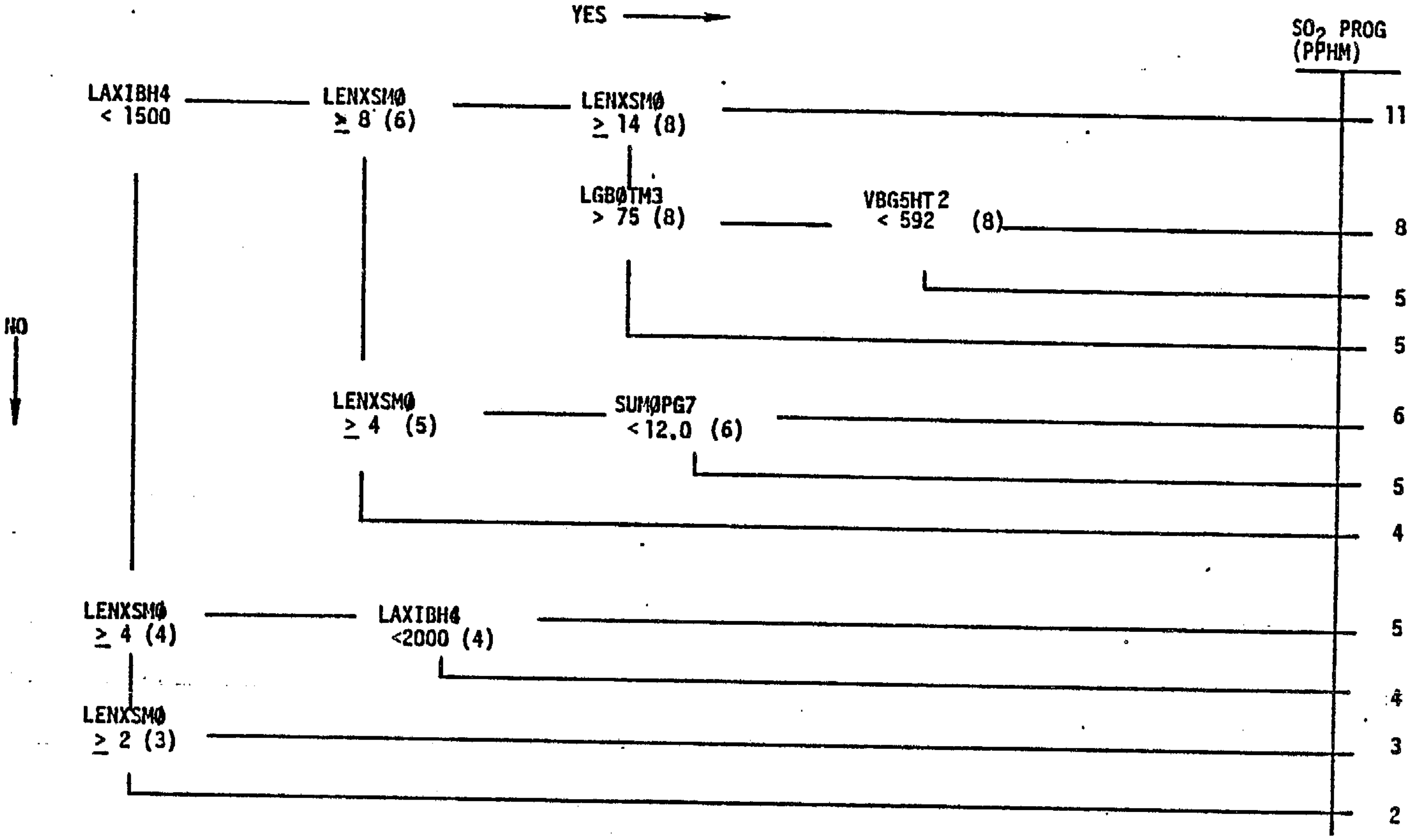
Figure 5.8 24-Hour Winter SO_2 Predictions for Lennox, $R^2 = 0.29$



(NOTE: Number encircle, ie (6), indicates predicted value if data are not available for further splits.)

Where: UPLAZM1 - Tomorrow's predicted 1-hr max ozone at UPLAND (PPHM)
LAXTPH7 - 7 a.m. pressure gradient between LAX -TPH (mb)
DAGØWV3 - 1 p.m. wind velocity at DAGGETT (mph)
LAXIDT4 - 7 a.m. LAX inversion top-bottom temperature (°C)
SDBØTM3 - 1 p.m. surface temp. at SDB (°F)
MFRWMC7 - 7 a.m. pressure gradient between MFR-WMC (mb)
SANWMC3 - 1 p.m. pressure gradient between SAN-WMC (mb)

Figure 5.9 24-Hour Summer SO_2 Predictions for Lennox, $R^2 = 0.40$



(NOTE: Number encircled, ie. (6), indicates predicted value if data are not available for further splits.)

Where: LAXIBH4 - 7 a.m. LAX inversion base height (Ft)

LENXSMØ - Today's 1-hr max SO_2 at Lennox (PPHM)

LGBØTM3 - 1 p.m. surface temp at LGB (°F)

VBG5HT2 - 500 mb height at VBG, 12 Z (10 m)

SUMØPG7 - Σ 7 a.m. pressure gradients: LAX-DAG, SAN-LAS, SDB-VCV (mb)

Figure 5.10 24-Hour Winter SO_2 Predictions for Fontana, $R^2 = 0.47$

5.3.2 Development of 24-Hour Algorithms

Decision tree and regression analyses were used to develop valid 24-hour SO_2 prediction algorithms. Due to changing meteorology and daily emission patterns, day-in-advance prediction is less sensitive than same-day algorithms to changes in SO_2 concentrations. As a result, the objective for the 24-hour forecast was merely to predict the potential for high SO_2 concentrations.

Fontana

The initial method used to develop a 24-hour forecast was stepwise multiple regression. First, the set of optimal linear predictors was defined from the correlation matrices. Several regression equations were formed using combinations of these variables. The most productive equation related Fontana (FONTSM1) with several afternoon variables: UPLAZM1, LAXTPH7, MFRWMC7, SDBØTM3. This equation had a correlation coefficient of $R = 0.52$. The resulting forecast predicted SO_2 poorly. Expanded regression analysis was then abandoned in favor of a decision tree approach to derive a working algorithm.

Several key linearly and nonlinearly related variables were selected for the ensuing decision tree analysis. Included were persistence and Upland's predicted oxidant value.* The decision tree attained a high prediction rating, explaining 66% of the variance in the SO_2 distribution, and was able to predict SO_2 values greater than 10 pphm.

The summer algorithm showed fair resolution in forecasting SO_2 values, outpredicting the 8-11 A.M. same-day decision tree. Differences between the same day and 24-hour algorithms are that the 8-11 A.M. model was designed to catch days when SO_2 is greater than or equal to 10 pphm in the morning (losing resolution for higher predictions) while the day-in-advance model was designed to predict higher values of SO_2 accurately (to count potential violation days).

*UPLAZM1 was used as the predicted value of tomorrow's oxidant concentration.

To further increase the resolution of the summer SO_2 forecasts the decision tree was used as a screen for a grouped regression analysis. This procedure was designed to segregate potential high concentration days and then produce a forecast based upon their relationships to a separate set of predictions. This combination of AID and regression is a form of a forced piecewise linear multiple regression. Since the addition of regression to AID in this case failed to improve the prediction resolution significantly, the 24-hour decision tree was only slightly modified (unnecessary splits were removed to simplify the forecast algorithm).

Lennox

Methodologies tested to produce a 24-hour prog for Lennox (both seasons) also included regression and decision tree analyses. Preliminary regression equations using the optimal set of predictors explained only 20.7% of the variance and could not accurately predict high SO_2 values.

SO_2 24-hour forecasting, using time series, was also included as a potential prediction algorithm. A 24-hour model was run and tested against the dependent data set. Because the algorithm suffered from a distinct time lag, never catching the start of a trend, it was abandoned in favor of decision tree analysis.

For both seasons, decision trees predicted high SO_2 values somewhat more accurately than regression. The winter decision tree explained 29.0% (497 cases) of the variance, more than the regression equation. The summer algorithm explained 42.7% (451 cases) of the variance, as opposed to 33.1% explained by the regression equation. Both decision tree (summer and winter) algorithms can forecast high SO_2 concentrations, $\geq$ 10 pphm.

Basin-Max

The same-day algorithm designed for the Basin-Max could not be adapted for a 24-hour prog. As a result only regression analysis was used to determine prediction algorithms for day-in-advance SO_2. Initial equations were determined from a limited set of morning and afternoon data, not including persistence. The resulting equations were fair (R^2 = .19, summer and R^2 = 0.24, winter), with a moderate ability to predict significant changes. With persistence included in the analysis, the amount of variance explained was R^2 = 0.24 for 470 summer cases and again R^2 = 0.24 for the winter.

The final equation set utilized one equation from each set of regression runs. The summer equation included persistence, LGBØTM3, and LAXIBH4, while the winter equation was based upon OAK5HT2, LAX8TMØ, and SUMØPG7.

5.3.3 Verification of the 24-Hour Algorithms

Table 5.3 gives the verification results for the dependent and independent data sets. As expected, the 24-hour forecasts of most models were less accurate than corresponding same-day predictions. One exception was the 24-hour summer prediction algorithm for Fontana, which displayed improved capabilities over its same-day algorithms.

While the algorithms for LENX and FONT showed improvement over persistence, the Basin-Max algorithms were only marginally better, especially on the independent test. This is a reflection of the localized problems in SO_2 conditions alluded to earlier (i.e. to pinpoint the highest SO_2 levels is more precise using site-specific algorithms than for the generalized case).

5.4 30-HOUR INITIAL ALGORITHMS

Thirty-hour forecast algorithms consisted of decision trees for Fontana and "perfect prog" regression equations for Lennox and the Basin-Max.

5.4.1 Final 30-Hour Algorithms

Presented in this section are the 30-hour forecast algorithms.

(1) Fontana

(see decision trees Figure 5.11 and 5.12)

Table 5.3 Overall 24-Hour SO_2 Prediction Rating

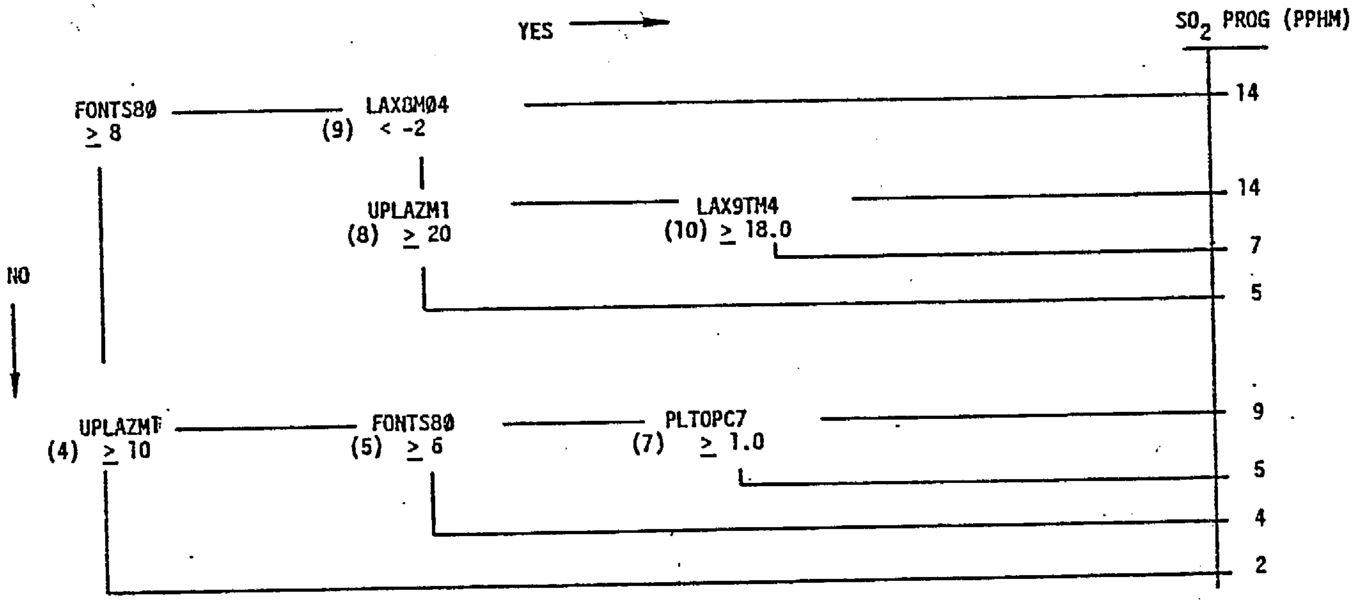
N	METHOD	T_c	- 10E	+ T_2	+ C	+ P	= R	R/PERFECT R
	PERFECT	100	0	100	100	25	325	
	DEPENDENT DATA SET: 1974-1976							
	MAY-OCT							
329	Fontana-persistence	84	26	61	0	25	144	.443
543	Lennox-persistence	91	19	71	0	25	168	.517
557	Basin-Max-persistence	71	26	62	0	25	132	.406
166	Fontana (algorithm)	91	18	84	48	25	230	.708
465	Lennox (algorithm)	93	15	82	34	25	219	.674
447	Basin-Max (algorithm)	74	23	67	18	25	161	.495
333	Fontana climatology	86	30	50	37	25	168	.517
549	Lennox climatology	93	20	70	40	25	208	.640
549	Lennox time series	93	17	78	32	25	211	.649
	NOV-APR							
296	Fontana-persistence	93	23	67	0	25	162	.498
527	Lennox-persistence	86	26	61	0	25	146	.449
541	Basin-Max-persistence	71	27	56	0	25	125	.385
225	Fontana (algorithm)	96	16	80	65	25	250	.769
497	Lennox (algorithm)	92	19	71	42	25	211	.649
403	Basin-Max (algorithm)	75	22	66	36	25	180	.554
304	Fontana climatology	96	22	68	48	25	215	.662
537	Lennox climatology	91	23	64	40	25	197	.606
537	Lennox time series	91	22	70	27	25	191	.588
	INDEPENDENT DATA SET: 1977							
	MAY-OCT							
141	Fontana-persistence	86	32	50	0	25	129	.397
182	Lennox-persistence	97	18	81	0	25	185	.569
183	Basin-Max-persistence	63	34	52	0	25	106	.326
141	Fontana (algorithm)	87	31	55	29	25	165	.508
182	Lennox (algorithm)	97	14	88	40	25	236	.726
177	Basin-Max (algorithm)	53	42	39	33	25	108	.332
	NOV-APR							
179	Lennox-persistence	89	17	73	0	25	170	.523
179	Basin-Max-persistence	82	26	62	0	25	143	.440
179	Lennox (algorithm)	89	26	73	38	25	203	.625
178	Basin-Max (algorithm)	73	29	50	31	25	150	.462

LEGEND

* = Best Method
N = Number of Predictions
T_C = Total Correct (%)
E = Mean Absolute Error (PPHM)

T_2 = Correct ±2 PPHM (%)
C = Significant Changes
Correct ±2 PPHM (%)

R = Rating
P = Score (Climatological
Constant)



(NOTE: Number encircled, ie, (9), indicates predicted value if data are not available for further splits.)

Where: FONTS8Ø - Today's predicted (8-11 a.m.) SO_2 at FONTANA (PPHM)

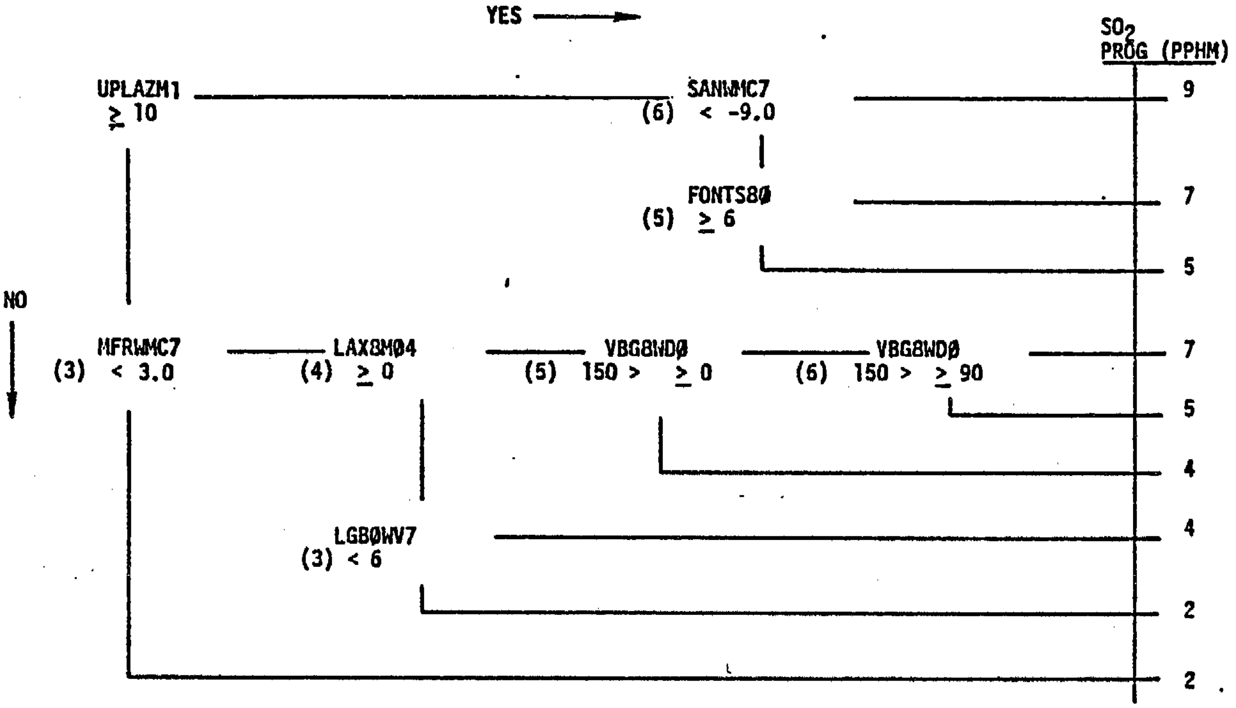
LAX8MØ4 - 7 a.m. 850 mb temp. - surface temp at LAX (°C)

UPLAZM1 - Tomorrow's predicted 1-hr max oxidant at UPLAND (PPHM)

LAX9TM4 - 7 a.m. 950 mb temp at LAX (°C)

PLTOPC7 - 7 a.m. average 24-hour pressure changes at WMC,RNO,TPH (mb)

Figure 5.11 30-Hour Summer SO_2 Prediction for Fontana, $R^2 = 0.66$



(NOTE: The number encircled, ie (6), indicates predicted value if data are not available for further splits.)

Where: UPLAZM1 - Tomorrow's predicted ozone at UPLA (PPHM)
SANWMC7 - 7 a.m. pressure gradient between SAN-WMC (mb)
MFRWMC7 - 7 a.m. pressure gradient between MFR-WMC (mb)
FONTS8Ø - Today's predicted (8-11 a.m.) SO_2 max at FONT (PPHM)
LAX8MØ4 - 7 a.m. 850 mb temp-surface temp at LAX (°C)
VBG8WDØ - 850 mb wind direction at VBG 00Z
LGBØWV7 - 7 a.m. surface wind velocity at LGB (MPH)

Figure 5.12 30-Hour Winter SO_2 Prediction for Fontana, $R^2 = 0.50$

(2) Lennox

Summer: LENXSM1 = -0.00045 LAXIBH4 + 0.12 VBG5HT1
+0.18 LENXSMY - 0.083 LAX8TM4
-0.047 SUMØPG7 - 62.1

Winter: LENXSM1 = 0.13 VBG5HT1 - 0.00016 LAX1BH4 - 59.6

(3) Basin Max

Summer: BASNSM1 = -0.0061 LAX1BH4 + 0.11 VBG5HT1
+0.12 BASNSMY + 0.12 VBG5HC1
- 0.058 VMW5HT1 - 53.8

Winter: BASNSM1 = 0.16 VBG5HT1 - 0.00026 LAX1BH4 - 84.2

5.4.2 Method Development

Three basic techniques were attempted to produce a 30-hour forecast: decision trees, stepwise multiple regression and the perfect prog.

The perfect prog method used for both Lennox and the Basin Max relates values of meteorological variables forecast for tomorrow by the numerical simulation models of the National Weather Service to values of SO_2. Thus, this SO_2 prediction relies upon accurate numerical meteorological forecasting.

The SO_2 prediction algorithms generally displayed poor 30-hour prediction accuracy, even worse than climatology. As a result, extensive 30-hour forecast algorithm development was limited. It is worth noting that the upper air progs used in the perfect prog procedure were not as effective as for oxidants. This indicates that large-scale features cannot effectively be used for long-term SO_2 prediction.

5.4.3 Validation of the 30-Hour Algorithms

The verification scores of both the dependent and independent data sets are given in Table 5.4. Note that climatology is a good predictor for the 30-hour prediction. Using the best algorithms, the degree of improvement over climatology is small. For Lennox,in fact, the verification of the new algorithm on the dependent data was not as good as climatology. Thus, it can reasonably be concluded that the nature of SO_2 build-up is such that

Table 5.4 Overall Prediction Rating for 30-hr Predictions (May - Oct)

Number of Predictions	Method	T_c	- 10E	+ T_2	+ C	+ P	= RATING	RATING ÷ PERFECT RATING
	PERFECT	100	0	100	100	25	= 325	
	Summer 1974 - 1976							
329	Fontana-persistence	84	26	61	0	25	144	.443
543	Lennox-persistence	91	19	71	0	25	168	.517
557	Basin-Max-persistence	73	28	55	0	25	125	.385
152	Fontana (algorithm)	91	19	81	42	25	220	.677
453	Lennox (algorithm)	78	22	55	22	25	158	.468
456	Basin-Max (algorithm)	76	25	59	28	25	163	.501
549	Climatology-Fontana	86	30	50	37	25	168	.517
549	Climatology-Lennox	93	20	70	40	25	208	.640
	Summer 1977							
141	Fontana-persistence	86	32	50	0	25	129	.397
137	Fontana (algorithm)	81	34	56	19	25	147	.452

the meteorological conditions necessary to predict daily concentration changes are identifiable up to 24-hours in advance. Beyond that time, climatology is the most convenient predictor.

5.5 LOS ALAMITOS CASE STUDY

The major goal of the Los Alamitos case study was to determine the effect of emissions on the ability to forecast ambient SO_2 levels. It was hypothesized that with both emissions and selected meteorological data, accurate prediction should be possible.

Data from the Haynes and Los Alamitos power plants were compiled for the three years of the analysis (1974-1976). The data consisted of the sum of the combined daily tonnage of SO_2 emissions emitted from the two power plants (Haynes-Alamitos SO_2 emissions today--HAALSEØ). Figures 5.13 through 5.15 show daily fluctuations in emissions for the study period.

Linear Regression of SO_2 (LSALSMØ) vs. HAALSEØ

To relate the effects of HAALSEØ same-day SO_2 levels monitored at Los Alamitos, a series of regressions was performed. The initial analysis, a simple linear regression data, showed no strong linear relationship between emissions and SO_2. For 477 summer cases, the correlation coefficient between the two variables was R = 0.33, with a standard error of 3.9 pphm. A graphical display of the basic relationship is shown in a scatterplot (Figure 5.16) where SO_2 data monitored at Los Alamitos are plotted versus HAALSEO. From the figure it can be seen that there exists a definite maximum potential for the formation of SO_2, given a distinct emissions level. Although there is a threshold potential for SO_2 formation, a number of different SO_2 concentrations can exist for each emissions level.

Multiple Regression of SO_2 vs. Meteorology

Stepwise multiple regression, using meteorology alone, produced an improved explanation of the existing SO_2 levels. Equations were developed for both seasons and were stratified by surface (LAXIBH4 = 110) and non-surface (LAX1BH4 > 110) inversions. Table 5.5 gives summary results of the equations, including the correlation coefficient, number of cases, percent variance explained, and the key variables in the equation.

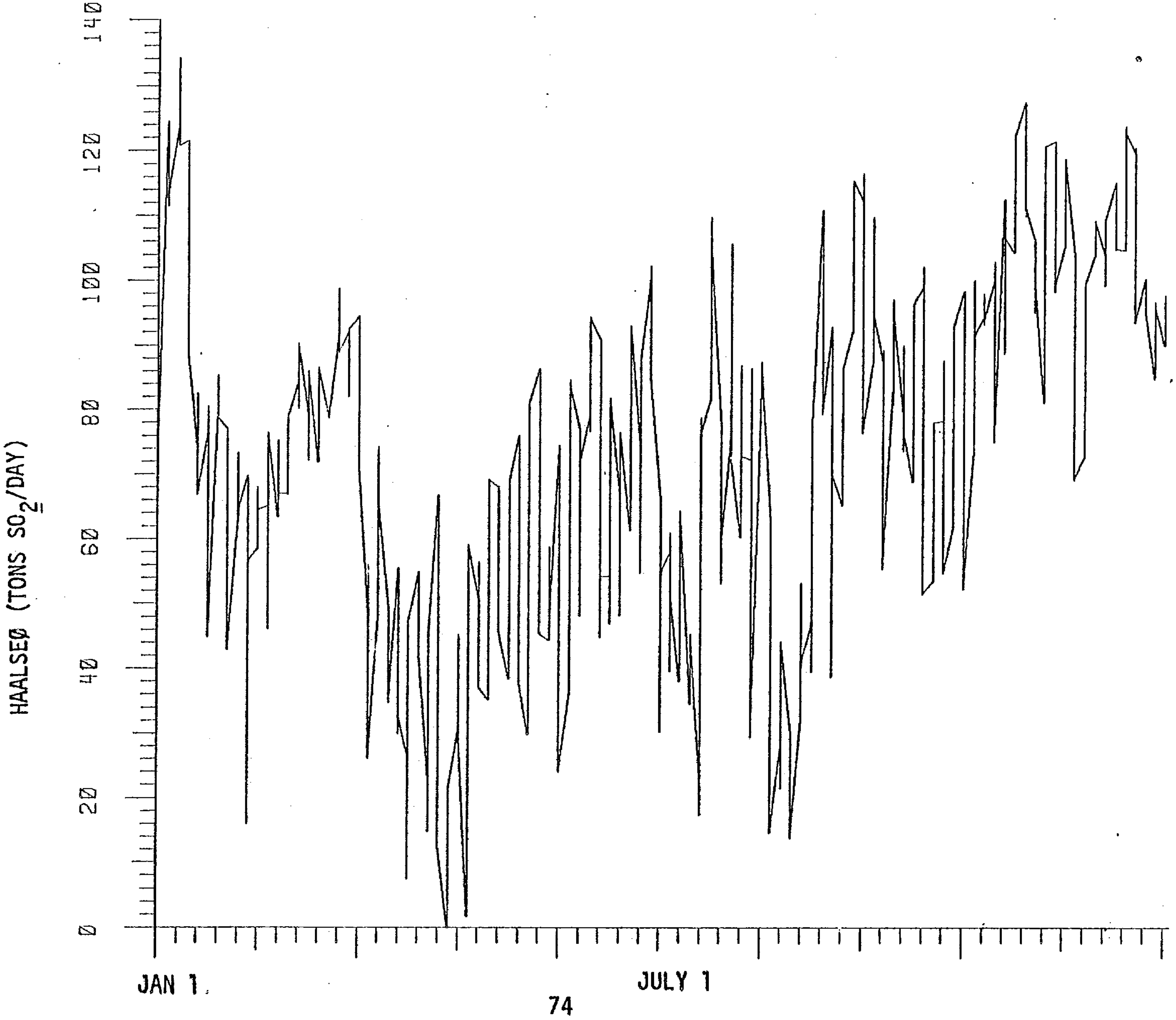


Figure 5.13 Daily Emissions of SO_2 From Haynes-Alamitos (HAALSEØ) Power Plants (TON-SO_2/Day) for 1974

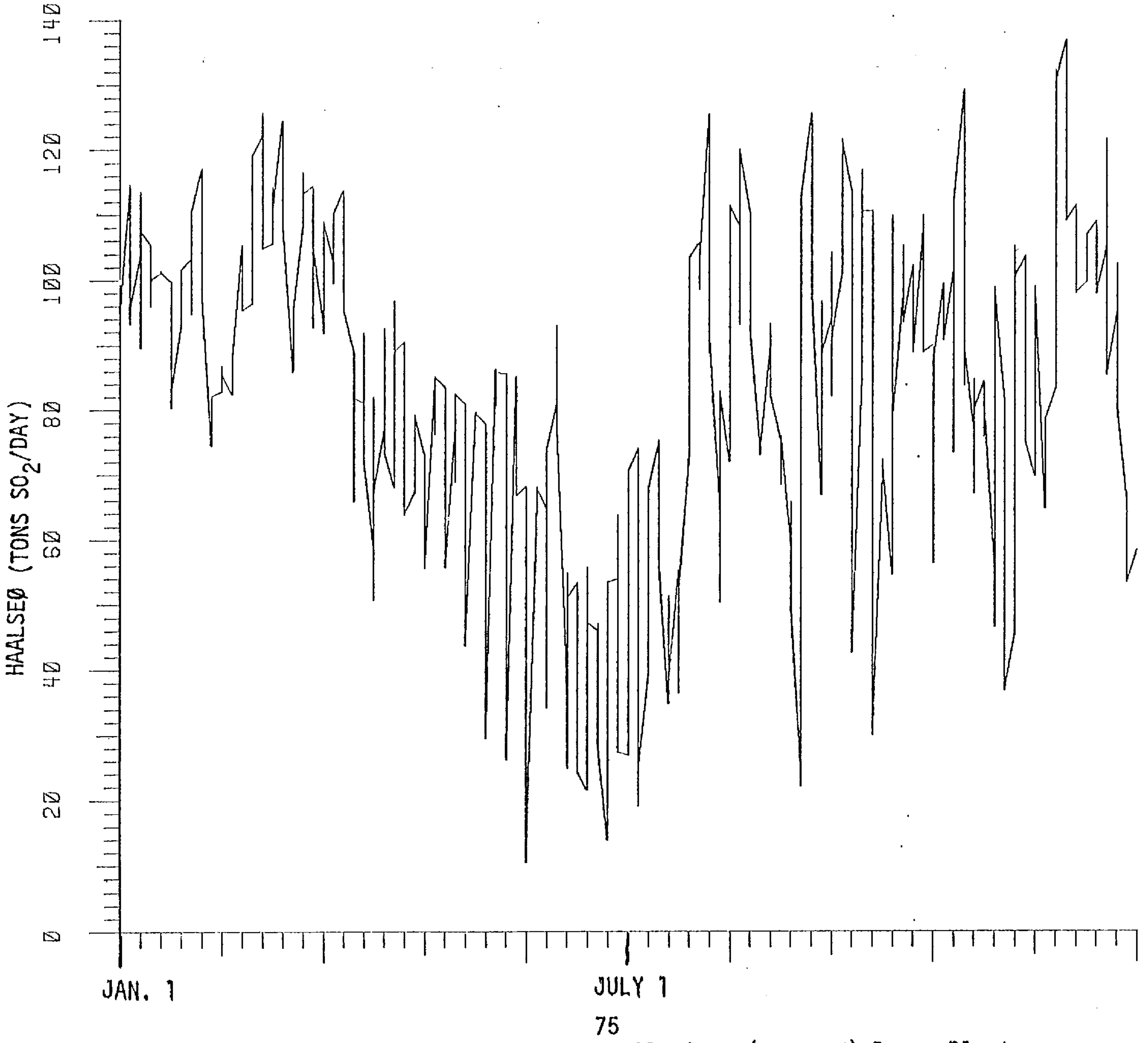


Figure 5.14 Daily Emissions of SO_2 From Haynes-Alamitos (HAALSEØ) Power Plants (TONS - SO_2/DAY) for 1975

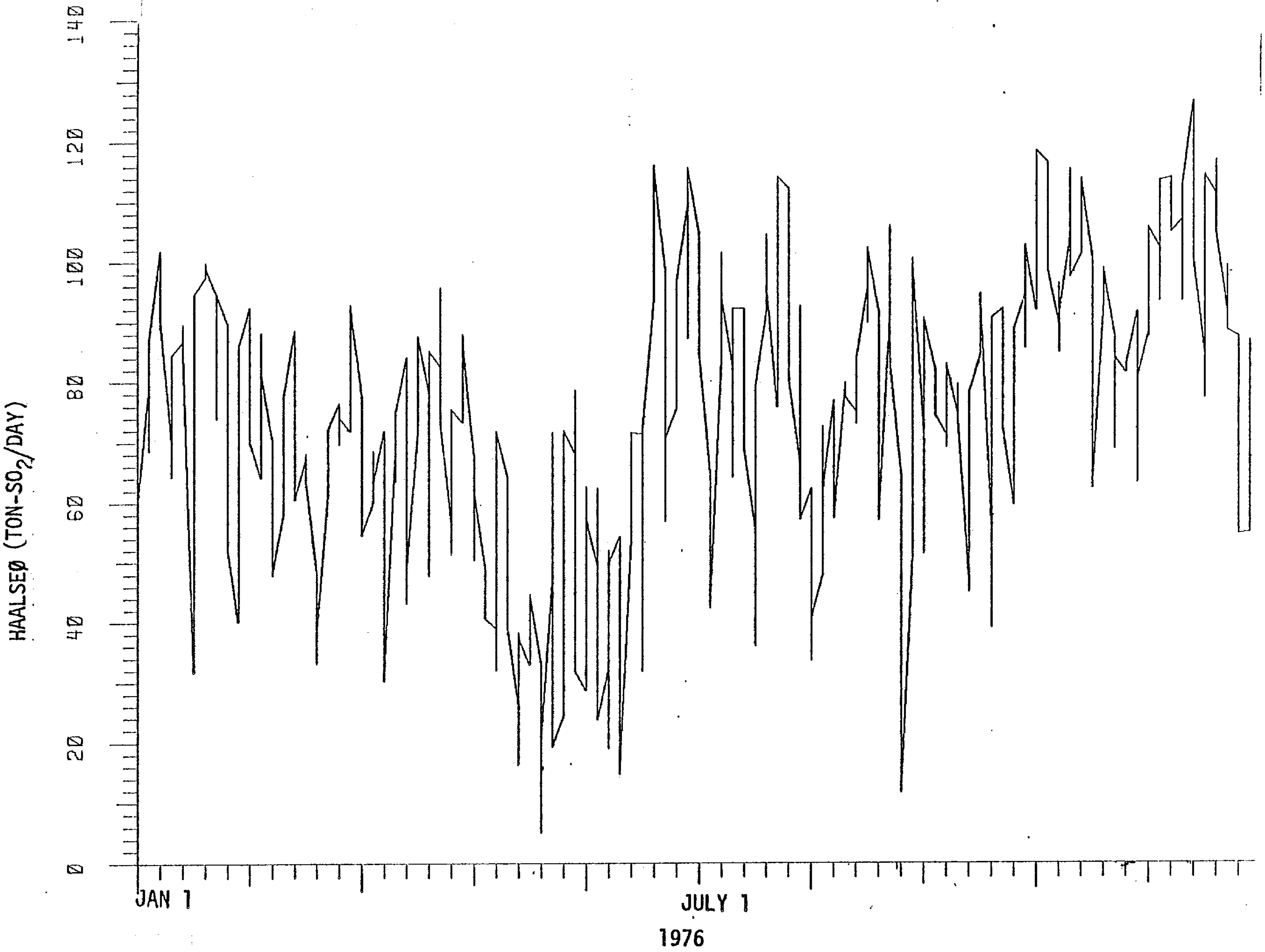


Figure 5.15 Daily Emissions of SO_2 From Haynes-Alamitos (HAALSEØ) Power Plants (Tons-SO_2/Day) for 1976

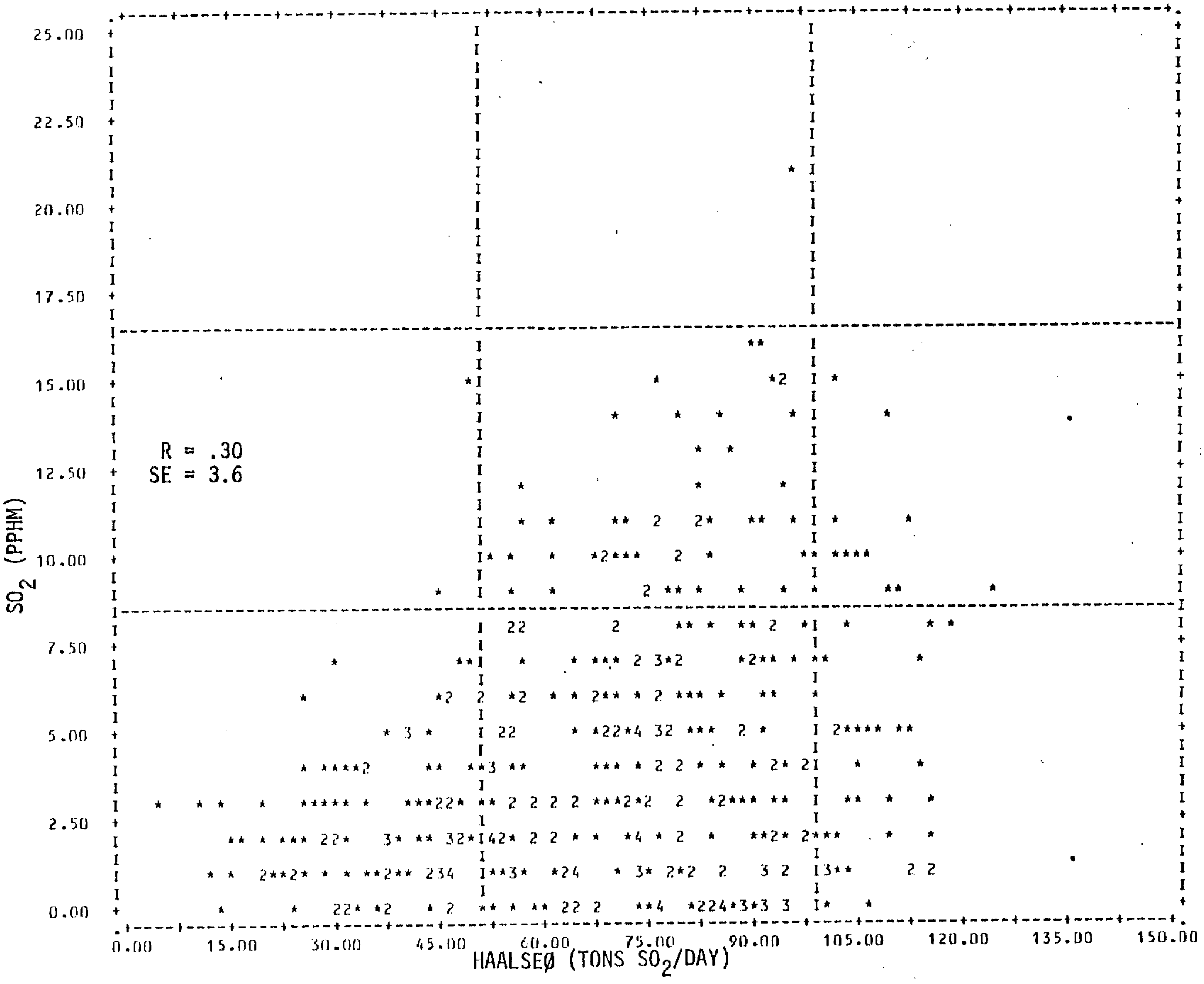


Figure 5.16 Scatterplot of HAALSEØ (TONS-SO_2/DAY) and LASALSMO (PPHM)

Table 5.5 Comparative Statistics for Regressions Predicting LSALSMO (Meteorology Only)

Season	Correlation Coefficient	Total Variance Explained	Standard Error (pphm)	Number of Cases	Key Variables	Variance Explained
Summer LAX1BH4 > 110	.57	32%	3.2	344	LAX1RHØ	16%
					LGBØVZ7	7%
					LAX9TM4	3%
					LAXØTM4	1%
					LGBØDP7	2%
					LAXIDT4	1%
					OTHERS	2%
Summer LAX1BH4 = 110	.74	55%	4.1	36	LAXTPH7	22%
					SUMØPG7	11%
					LAX9TM4	5%
					LGBØDP7	5%
					LAX8MØ4	3%
					DSRTPC7	3%
					LAX1RHØ	1%
					OTHERS	5%
Winter LAX1BH4 > 110	.54	29%	3:5	151	LAXØWV7	13%
					SUMØPG7	5%
					SAN7RH2	3%
					LGBØVZ7	2%
					LAXØTM4	1%
					LAX1BH4	1%
					OTHERS	4%
Winter LAX1BH4 = 110	.44	20%	3.1	166	LAXØWV7	5%
					SUMØPG7	3%
					LGBØVZ7	3%
					LAX1RHØ	2%
					SAN7RH2	1%
					LAXØTM4	1%
					LGBØDP7	2%
					SAN5HT2	1%
					OTHERS	2%

Separate regressions were generated for days with and without surface-based inversions, and for the two seasons. For the summer surface inversion category (5% of all cases), the regression equation accurately predicted the higher values of SO_2. The equations for the 3 remaining categories (95% of all cases), with correlation coefficients ranging from 0.42 to 0.55, were not accurate enough to use as predictors.

Multiple Regression of SO_2 vs. Emissions and Meteorology

To combine the effects of emissions and meteorology, additional regressions were generated using both the original data set and HAALSEØ. Summer regressions were developed for each of the categories (see Table 5.6). The new equation set predicted only slightly better than the equations using meteorology alone.

HAALSEØ placed the third best predictor for the summer non-surface-based inversion equation, explaining an additional 4% of the total variance. In the surface-based inversion set HAALSEO contributed only an insignificant 0.17% to the total variance explained.

For the winter regressions (approximately 48% of the total cases) HAALSEØ was not flagged as a potential predictor. In both algorithms, HAALSEØ contributed less than 1% to the total variance explained.

Scatterplot Analysis

Summer season scatterplots of several meteorological variables suggested that many of the potential predictors were not linearly related to SO_2. (See Figures 5.17 through 5.21.) Also shown are the individual correlation coefficients and the standard error. The large standard error and the scatterplots themselves illustrate the large variance in the SO_2 distribution compared to the best fit regression line. In all of the scatterplots the higher values of SO_2 ($\geq$ 10 pphm) are not adequately represented by the straight line fit. Also, in many of the figures, the actual regression best fit is not clearly shown by the shape of the SO_2 distribution.

Table 5.6 Comparative Statistics for Summer Regressions Predicting LSALSMO (Meteorology and Emissions)

Condition	Correlation Coefficient	Total Variance Explained	Standard Error	Number of Cases	Key Variables	Variance Explained
Summer LAX1BH4 > 110	.60	36%	3.1	344	LAX1RHØ	16%
					LGBOVZ7	7%
					HAALSEØ	4%
					SUMØPG7	2%
					LAXØTM4	1%
					LGBØDP7	2%
					DSRTPC7	1%
					OTHERS	3%
Summer LAX1BH4 = 110	.74	55%	4.2	36	LAXTPH7	22%
					SUMØPG7	11%
					LAX9TM4	5%
					LGBØDP7	5%
					LAX8MØ4	3%
					DSRTPC7	3%
					LAX1RHØ	1%
					OTHERS	5%

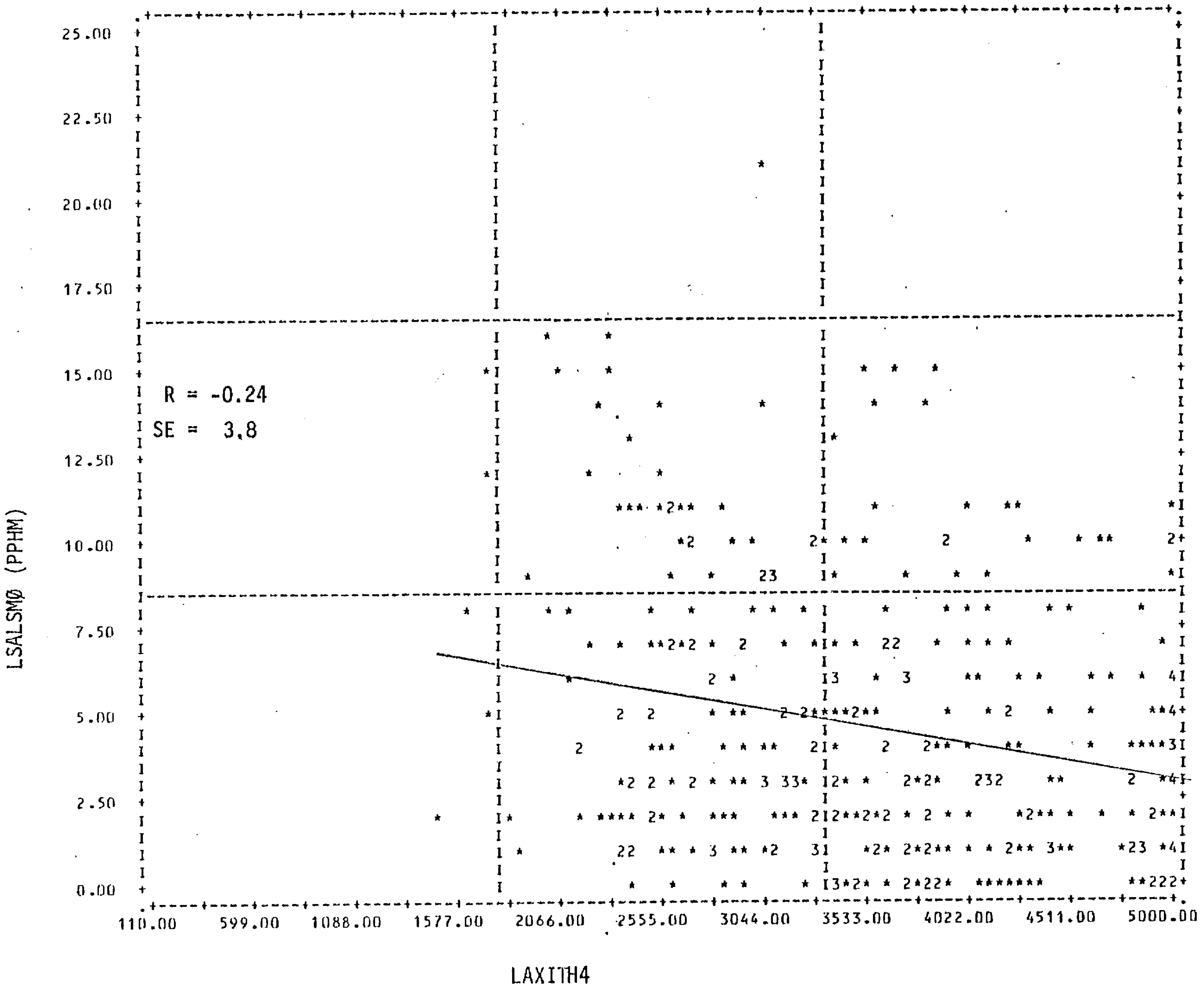


Figure 5.17 Scatterplot of LSALSMØ (PPHM) and LAX Morning Inversion Top Height LAXITH4, (Ft); Summer.

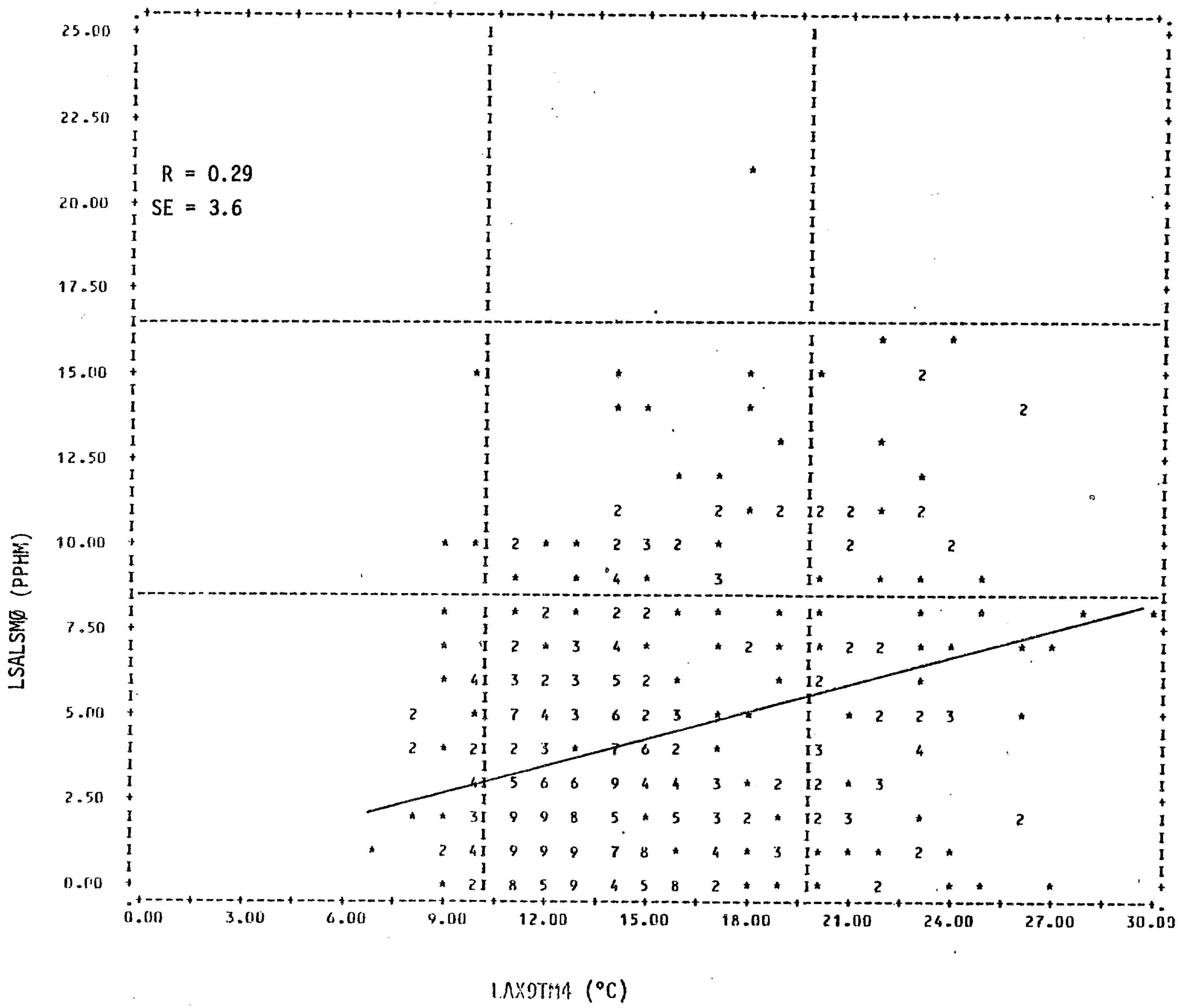


Figure 5.18 Scatterplot of LSALSMØ (PPHM) and morning LAX 950 mb Temp LAX9TM4 (°C); summer

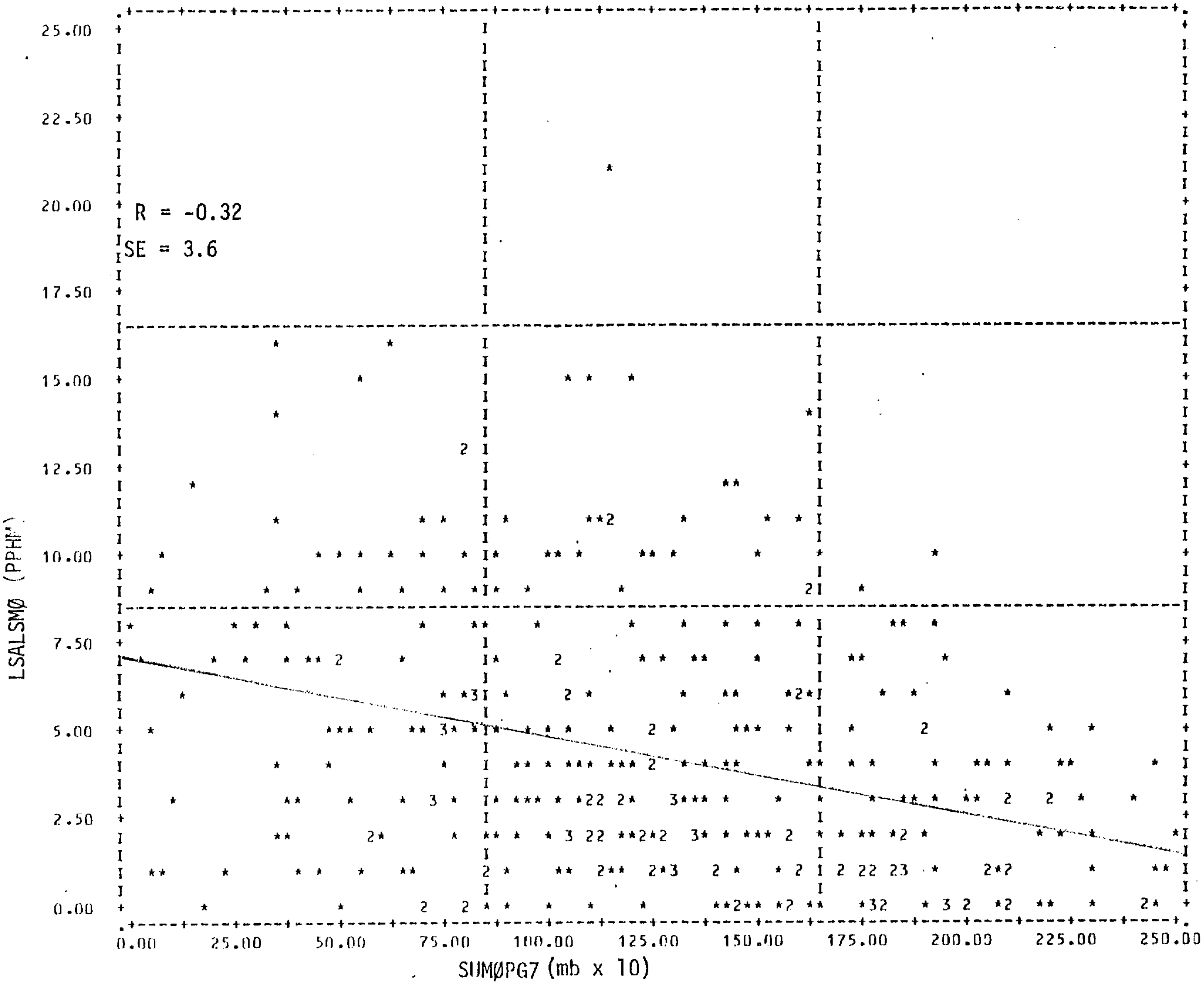


Figure 5.19 Scatterplot of LSALSMØ (PPHM) and SUMOPG7 (mb x 10)

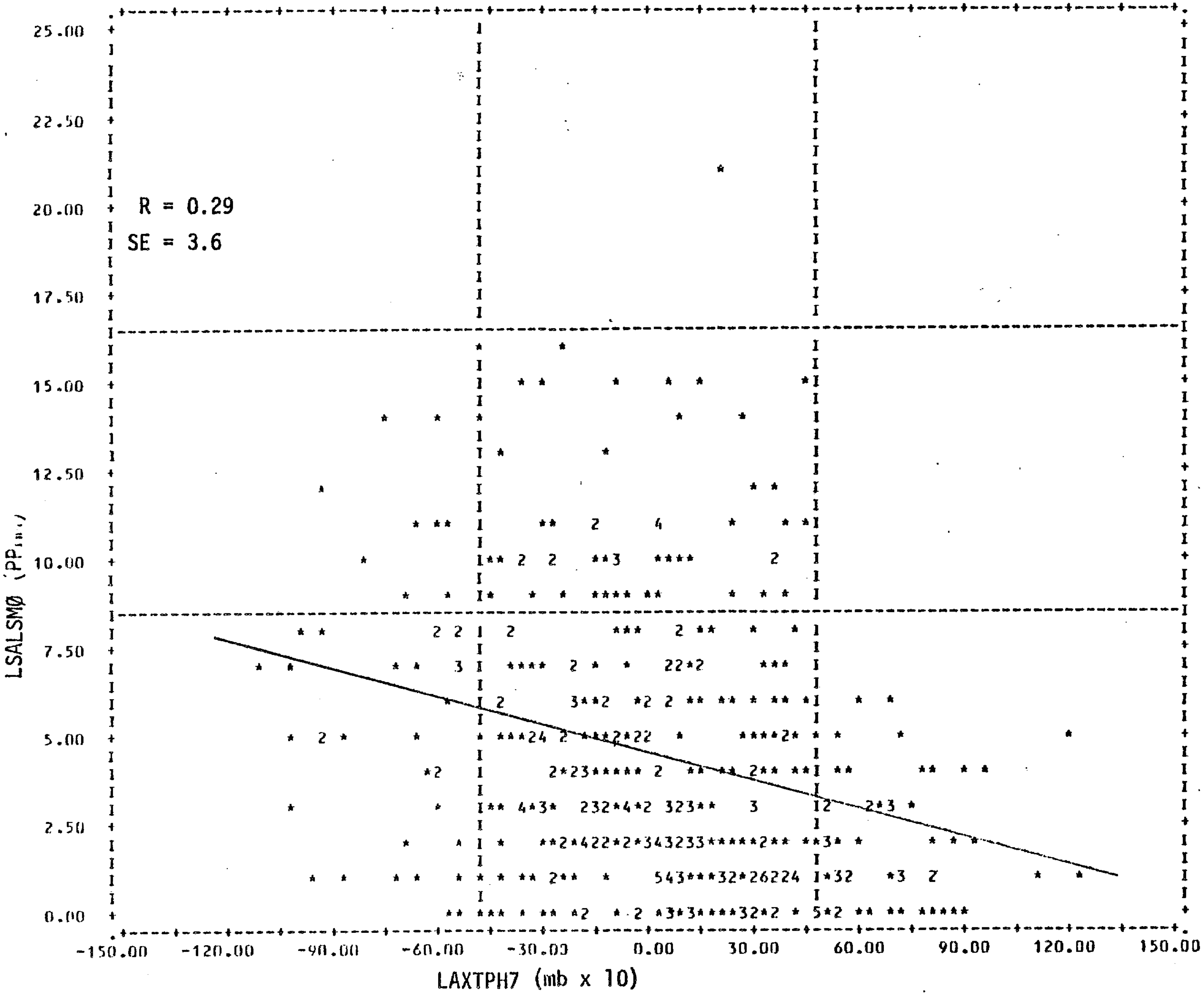


Figure 5.20 Scatterplot of LSALSMØ (PPHM) and LAX-TPH 7AM Pressure Gradient LAXTPH7 (mb x 10); Summer

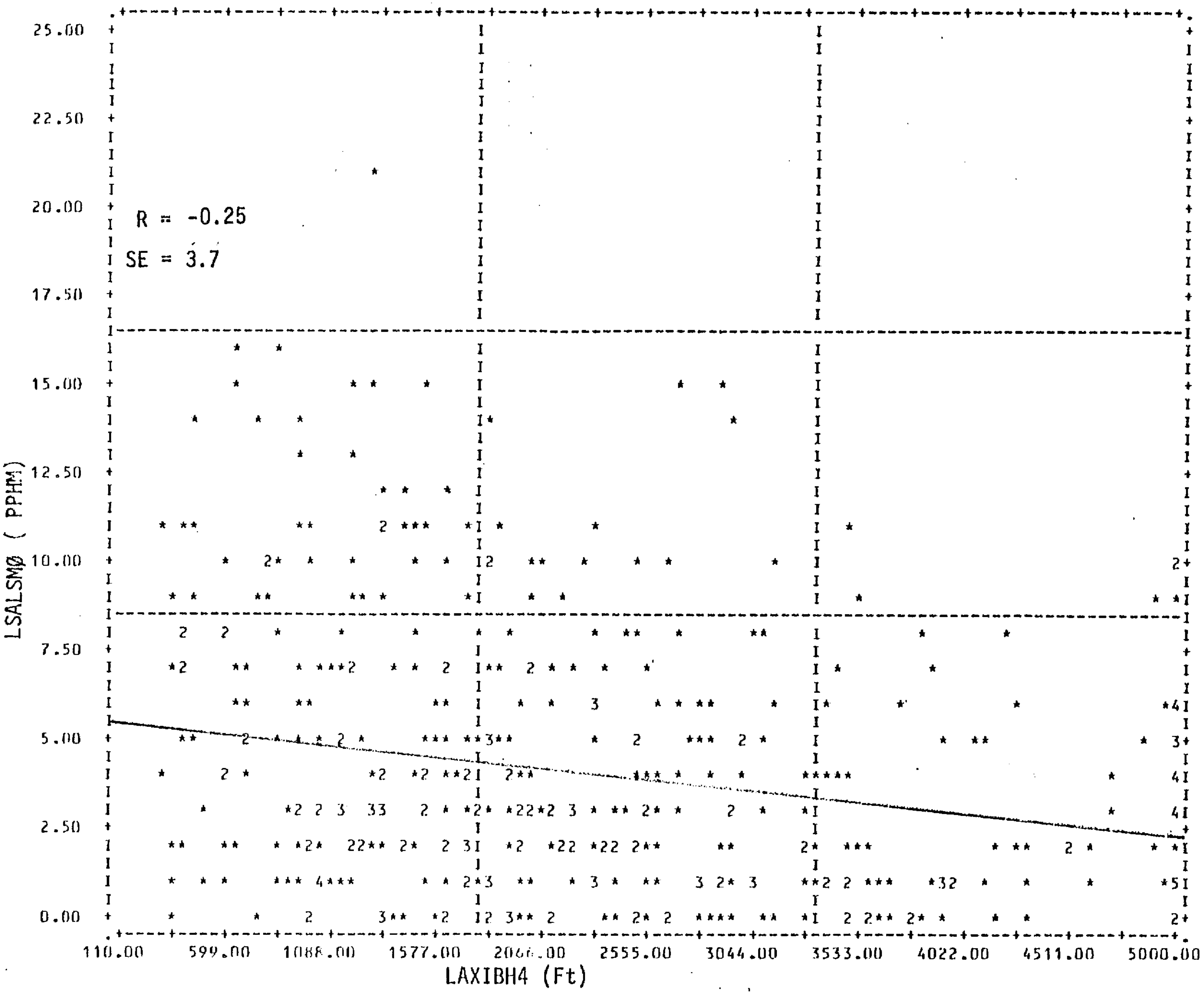


Figure 5,21 Scatterplot of LSALSMØ (PPHM) and LAX Morning Inversion Base Height LAXIBH4, (Ft); Summer

Decision Tree Analysis

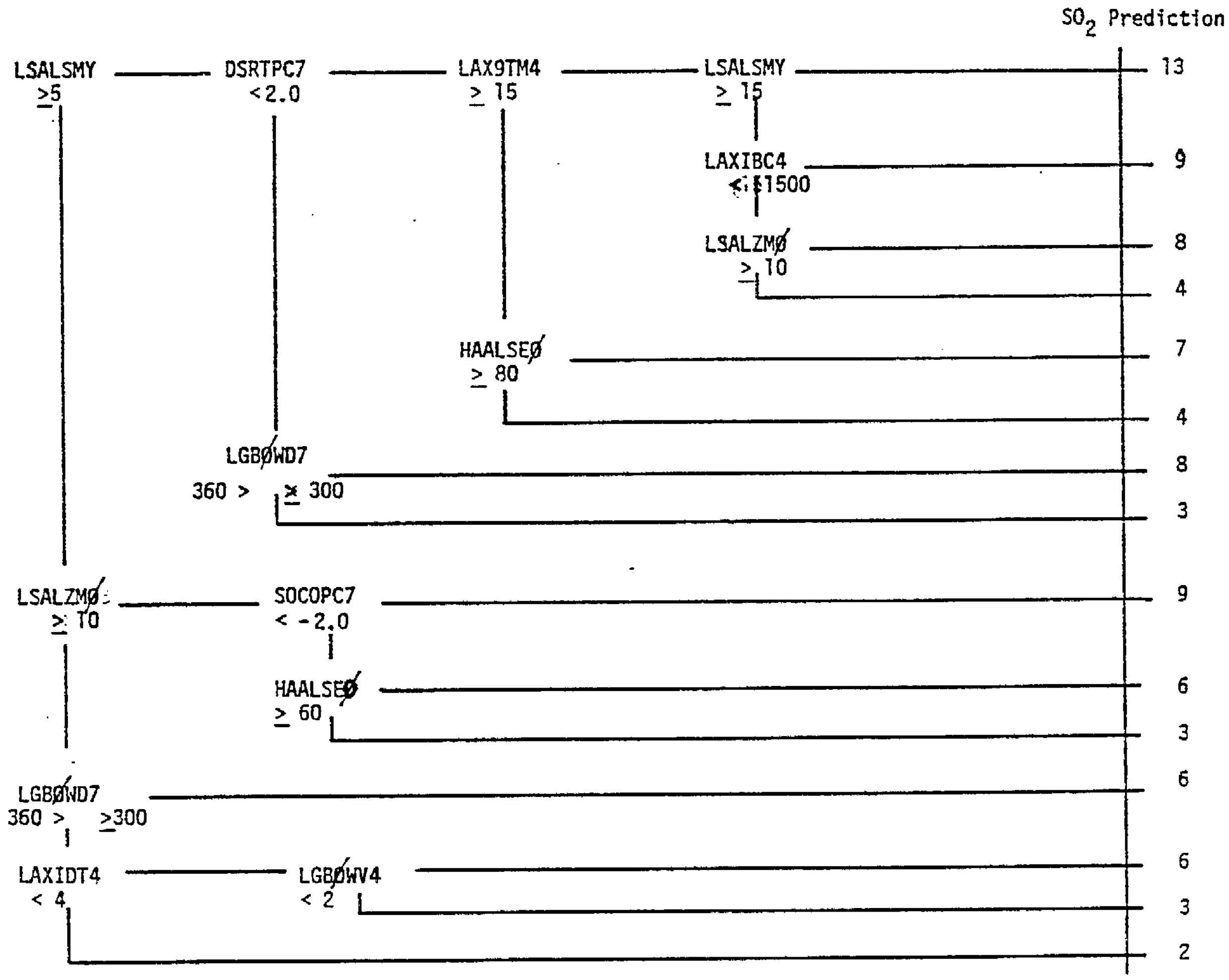
To account for contributions to a prediction algorithm from non-linearly related variables, especially wind direction, decision tree analysis was implemented. The use of the decision tree allowed the inclusion of local wind parameters for LGB and LAX (7 AM and 4 AM) as well as upper air wind flow for Vandenburg.

The new set of predictors included persistence, wind directions and emissions data to complement the basic set of meteorological variables. Two different AID algorithms were produced; summer, with LAX1BH4 ≠ 110 including the entire variable set, and winter using just meteorology alone. The summer LSAL decision tree (Figure 5.22) had a correlation coefficient of 0.70 (330 cases) with HAALSEØ emissions explaining a total of 3.4% of the variance (both splits). The role of "in hand" emissions data in prediction was far overshadowed by the contributions of persistence, explaining 14.5% of the variance. The remaining 30.6% of the variance was explained by meteorology and oxidant forecasts. The importance of persistence compared to emissions is evident through the splitting order, where the decision tree split 1st and 6th on persistence and 12th and 13th on HAALSEØ.

To estimate the potential of a predictive algorithm not including emissions data, the winter decision tree for Los Alamitos was developed without either SO_2 emissions or persistence. (See Figure 5.23.) Using meteorological variables only the tree achieved a correlation coefficient of .66 for all (310) winter cases. Without emissions and persistence the tree predicted more accurately than the winter regression equation including both of these variables. (The regression equation with all base heights achieved a correlation coefficient of .50.)

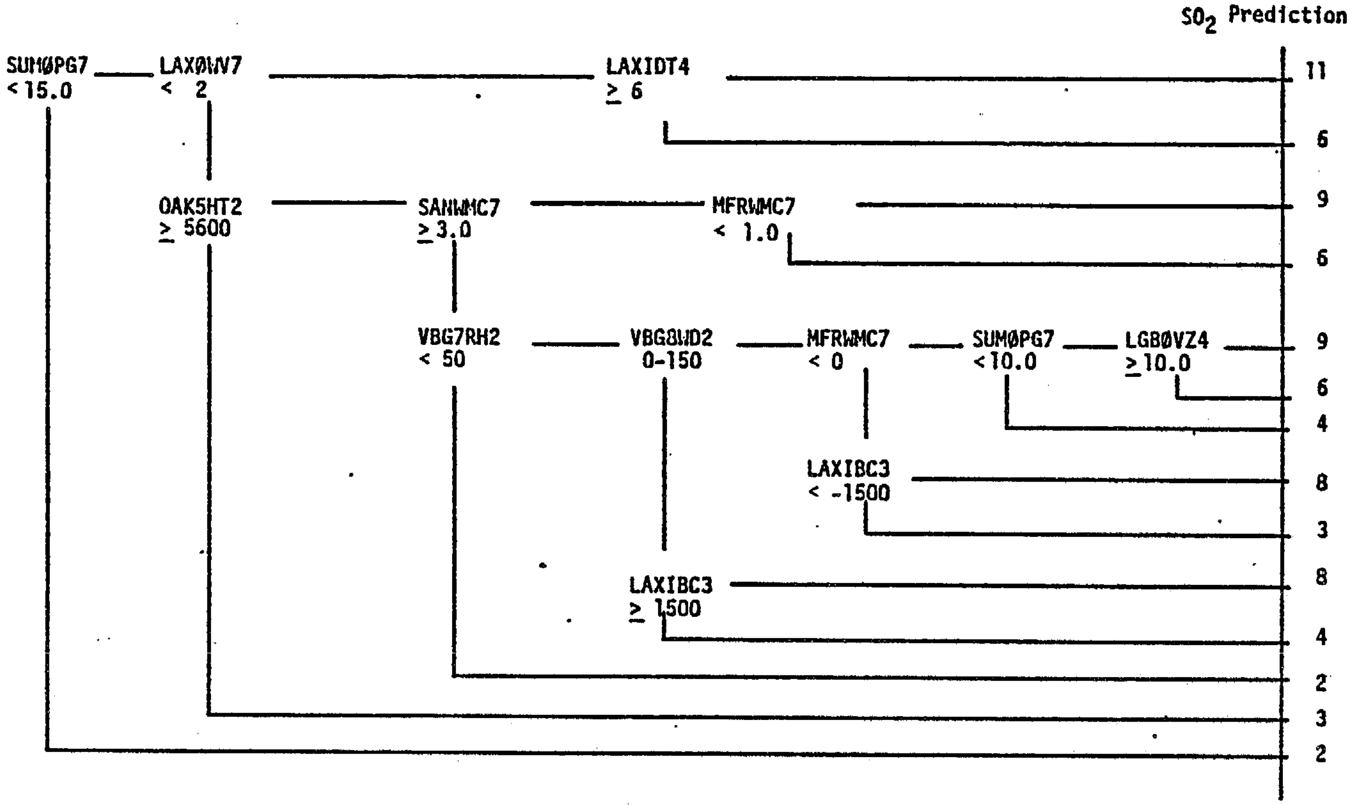
Summary

From this series of analyses several conclusions can be made. The amount of emissions establishes a ceiling for potential SO_2 build-up for LSAL. SO_2 prediction based solely upon the input of meteorology explained a substantial percentage of the variance in the observed SO_2 at Los Alamitos. The inclusion of emissions data did not necessarily improve forecast capabilities significantly enough to justify the need for a daily emissions



Where: LSALSMY - One hour max SO_2 at LSAL yesterday (PPHM)
DSRTPC7 - Avg. 7 a.m. 24-hour pressure changes at Σ DAG, LAS, TRM, YUM (mb)
LAX9TM4 - 7 a.m. 950 mb Temp at LAX (°C)
LAXIBC4 - 7 a.m. inersion thickness at LAX (ft)
LSALZMØ - Today's 1-hour max oxidant at LSAL (PPHM)
HAALSEØ - Today's Haynes and LSAL SO_2 emissions (tons/day)
LGBØWD7 - 7 a.m. surface wind direction at LGB
SOCOPC7 - Avg. 7 a.m. 24-hour pressure changes at Σ LAX, SAN (mb)
LGBØWV4 - 4 a.m. surface wind speed at LGB (MPH)

Figure 5.22 Same Day Summer SO_2 Decision Tree Prediction Algorithm for Los Alamitos



Where:

SUMØPG7	-	Σ 7 a.m. pressure gradients: LAX-DAG, SAN-LAS, SDB-VCV (mb)
LAXØWV7	-	7 a.m. surface wind velocity at LAX (MPH)
LAXIDT4	-	7 a.m. LAX inversion Top Temp - Base Temp (°C)
OAK5HT2	-	500 mb heights at OAK 12Z (10 m)
SANWMC7	-	7 a.m. pressure gradient between SAN-WMC (mb)
MFRWMC7	-	7 a.m. pressure gradient between MFR-WMC (mb)
VBG7RH2	-	700 mb relative humidity at VBG 12 Z (mb)
VBG8WD2	-	850 mb wind direction at VBG 12 Z (mb)
LGBØVZ4	-	4 a.m. surface visibility at LGB (miles)
LAXIBC3	-	7 a.m. LAX 24-hr. inversion base height change (ft)

Figure 5.23 Same Day Winter SO_2 Decision Tree Prediction Algorithm for Los Alamitos

projection. In the summer, emissions do play a limited role in the prediction algorithm, yet the inclusion of emissions only replaces one potential predictor with another. When included, persistence was usually a dominant factor in the algorithms. It is possible that persistence and emissions are so highly correlated that there is no advantage to including both in the analysis.

CHAPTER 6

ESTIMATES FOR MISSING DATA

In order for prediction algorithms to be utilized, input data must be available. On any given day, key meteorological data may not be available due to several factors:

(1) instrumentation problems at the originating site

(2) communications problems (teletype or facsimile outages, line garbling, etc)

When such conditions occur, it is especially important to estimate the input variables so that the predictions of pollutant concentrations can still be obtained objectively. Due to the large number of variables used in the complete set of prediction algorithms, and due to the seasonality factors involved in many of the meteorological parameters, accurate statistical prediction of these variables is not possible. Instead we have approached this problem in two ways: (1) a climatological summary, and (2) a method to estimate key LAX inversion variables. The following subsections describe these approaches.

6.1 Climatological Summaries

For each of the variables in the data base, descriptive statistics were compiled by month, indicating the mean and standard deviation, the maximum and minimum values, and characteristics of the distribution (skewness and kurtosis). Data used were for the 1974-6 period. Tabular listings are given in Appendix D.

In many cases, parameters can be accurately estimated subjectively. For example, the DOLA 24-hour algorithm uses the VBG 500 mb height. The exact height value at VBG may be missing from the NMC 500 mb analysis, but interpolation from other available sites (i.e. OAK and SAN) can give an accurate estimate of the needed parameter. Similarly, a teletype outage may cause the loss of a scan of surface observations; however, an estimate of the needed parameter, based on the previous hours data, can be accomplished quickly and accurately.

More significant problems generally occur if missing data are from upper air parameters (not taken every hour) and with no means of interpolation. Also, estimates tend to be more inaccurate as the number of hours increase between the last observed value and needed value. In these cases, the estimates can be guided by the climatological data. Again, using the 24-hour DOLA algorithm as an example, one of the needed parameters is the SAN-LAS pressure gradient at 21Z (SANLAS3). Let us assume that the last available was at 15Z (SANLAS7). Determine the departure from the monthly normal for SANLAS7, and then apply that correction factor to the monthly average listed for SANLAS3. This technique will allow for the duirnal effects that occur in surface pressure parameters.

It is important to realize that these techniques are intended to be a constructive guide for estimating key parameters and that thoughtful subjective applications are likely to improve the results.

6.2 Estimating the 14Z LAX Sounding from the El Monte Sounding

Many of the key meteorological input variables are taken from the LAX inversion data. Thus it is important to accurately estimate these parameters when such data are not readily available. In such instances, it is convenient to obtain a best estimate of the LAX inversion profile. Using the nearby El Monte (EMT) 14Z RAOB data for the years 1974-1976, multiple regressions were run to estimate the following 14Z LAX atmospheric variables:

(1) Inversion Base Height (FT)
(2) Inversion Top Height (FT)
(3) Inversion Base Temperature (°C)
(4) Inversion Top Temperature (°C)
(5) Surface Temperature (°C)
(6) 950 mb Temperature (°C)
(7) 850 mb Temperature (°C)

Initially, the data were grouped into three seasonal subsets: March-May, June-October, and November-February. For each subset, regression equations were generated for each of the seven key variables listed previously. Next, the data were re-grouped according to the El Monte 850 mb temperature: < 14.0°C, 14.0-21.9°C, and ≥ 22.0°C. Again, equations were generated for the seven RAOB variables. A matrix was constructed for each possible combination of the two subsets, and in each case, the regression equation which produced the least standard error was selected. Equations with the greatest standard error were rejected.

The remaining equations are tabulated in Table 6.1 with the corresponding standard errors. For convenience, the matrix in Table 6.2 indicated the appropriate set of equations for each combination of subset criteria.

From Tables 6.1 and 6.2, one can thus construct a reasonable estimate of the LAX inversion profile (RAOB) using only the available EMT RAOB data.

Table 6.1 LAX RAOB Predictive Equations

EQUATION #		S_e
	BASE HEIGHT	
(1)	$H_B = .48\ h_B - 72.4\ t_9 + 2093$	$\pm$ 345
(2)	$H_B = .87\ h_B + 57.2\ t_s - 560$	$\pm$ 562
(3)	$H_B = .55\ h_B - 94\ t_T + 2922$	$\pm$ 831
(4)	$H_B = .16\ h_B + 263\ t_s - 236\ t_B + 1332$	$\pm$ 1250
	TOP HEIGHT	
(5)	$H_T\ .28\ h_T - 84.5\ t_9 + 4028$	$\pm$ 595
(6)	$H_T = .41\ h_B + .47\ h_T + 972$	$\pm$ 770
(7)	$H_T = 250\ t_s - 148\ t_T + 2609$	$\pm$ 1365
	BASE TEMP	
(8)	$T_B = .61\ t_B + .15\ t_9 + 2.7$	$\pm$ 1.38
(9)	$T_B = .44\ t_B - .0005\ h_T + .28\ t_9 + 4.5$	$\pm$ 1.54
(10)	$T_B = .53\ t_B + .11\ t_T - .16\ t_s + .24\ t_9 + 2.3$	$\pm$ 1.75
(11)	$T_B = .37\ t_T - .2\ t_s + .43\ t_9 + 0.8$	$\pm$ 2.20
(12)	$T_B = .33\ t_T + .37\ t_8 + 1.2$	$\pm$ 2.45
	TOP TEMP	
(13)	$T_T = .95\ t_T + 1.1$	$\pm$ 1.21
(14)	$T_T = .0005\ h_T + .97\ t_T - 1.0$	$\pm$ 1.71
(15)	$T_T = .5\ t_T + .58\ t_8 + 1.3$	$\pm$ 2.36
(16)	$T_T = .0008\ h_T + .83\ t_T + .31\ t_9 - 3.9$	$\pm$ 2.47

Table 6.1 LAX RAOB Predictive Equations (Continued)

EQUATION #		S_e
	SURFACE TEMP	
(17)	$T_S = .5\ t_s + .08\ t_9 + 7.6$	$\pm$ 1.15
(18)	$T_S = .55\ t_s + .15\ t_T + 4.9$	$\pm$ 1.42
	950 mb TEMP	
(19)	$T_9 = .94\ t_9 + 0.7$	$\pm$ 1.33
(20)	$T_9 = .77\ t_T + .16\ t_9 + .0007\ h_B + 0.1$	$\pm$ 1.63
	850 mb TEMP	
(21)	$T_8 = .95\ t_B + 0.7$	$\pm$ 1.39
(22)	$T_8 = .001\ h_T + .78\ t_T - .17\ t_8 + 3.9$	$\pm$ 1.33

LEGEND:

EMT, LAX

h_B , H_B = HEIGHT OF INVERSION BASE(FT)

h_T , H_T = HEIGHT OF INVERSION TOP (FT)

t_B , T_B = TEMPERATURE OF INVERSION BASE (°C)

t_T, T_T = TEMPERATURE OF INVERSION TOP (°C)

t_s , T_s = SURFACE TEMPERATURE (°C)

t_9 , T_9 = 950 mb TEMPERATURE (°C)

t_8 , T_8 = 850 mb TEMPERATURE (°C)

Table 6.2 Matrix for Determining Appropriate Equations for Estimating LAX 14Z RAOB Variables

EL MONTE 850 mb TEMP (°C)

Season	< 14.0		14.0 - 21.9		≥ 22.0	
	VAR	EQ	VAR	EQ	VAR	EQ
Mar-May	H_B -	3	H_B -	2	H_B -	1
	H_T -	7	H_T -	6	H_T -	5
	T_B -	11	T_B -	9	T_B -	8
	T_T -	16	T_T -	13	T_T -	13
	T_s -	18	T_s -	18	T_s -	18
	T_9 -	19	T_9 -	19	T_9 -	19
	T_8 -	21	T_8 -	21	T_8 -	22
	VAR	EQ	VAR	EQ	VAR	EQ
Jun-Oct	H_B -	4	H_B -	2	H_B -	1
	H_T -	7	H_T -	6	H_T -	5
	T_B -	10	T_B -	9	T_B -	8
	T_T -	14	T_T -	13	T_T -	13
	T_s -	17	T_s -	17	T_s -	17
	T_9 -	19	T_9 -	19	T_9 -	19
	T_8 -	21	T_8 -	21	T_8 -	21

Table 6.2 Matrix for Determining Appropriate Equations for Estimating LAX 14Z RAOB Variables (Continued)

	< 14.0		14.0 - 21.9		≥ 22.0	
	VAR	EQ	VAR	EQ	VAR	EQ
Nov-Feb	H_B -	4	H_B -	2	H_B -	1
	H_T -	7	H_T -	6	H_T -	5
	T_B -	12	T_B -	9	T_B -	8
	T_T -	15	T_T -	13	T_T -	13
	T_s -	18	T_s -	18	T_s -	18
	T_9 -	20	T_9 -	19	T_9 -	19
	T_8 -	21	T_8 -	21	T_8 -	21

CHAPTER 7

LONG-TERM TREND REVISION

Prediction algorithms developed from a three-year base period (1974-1976) maximize the relationships between meteorology and pollutant concentrations. Since daily emission values are virtually impossible to obtain on a real-time basis, it was necessary to exclude emissions from the development of the prediction models. One assumes, therefore, that emissions have remained relatively constant over the base period.

As emission control programs affect the balance of emissions of both primary and photochemical precursor pollutants, the direct relationship between meteorology and observed ambient pollution concentrations can change. This can lead to systematic overpredictions at some future time (if pollutant trends are downward) or underpredictions(if trends are upward). Rather than reconstruct new algorithms at some future time with a revised data base, it is more feasible to develop methods of adjusting existing algorithms to detect and correct trend biases. Such methods should also be relatively simple to apply, so that daily real-time prediction can remain an efficient procedure.

The purpose of this chapter, therefore, is to provide two simple methods for the detection and correction of the systematic bias which may be present in the prediction equations. The central idea of the methods is to treat the prediction process as a black box, and make the correction solely based upon past input (observed values) and output (predicted values). The methods are presented in the form of a user's manual and are applied to a test data set for illustration purposes. Finally, the two methods are tested on two sets of real data for a comparison and are shown to be effective in reducing the total errors of prediction.

In order to first detect a trend, usually one has to analyze the data for an entire year and compare results to preceding years. For algorithm correction, however, it is not appropriate to wait that long, since adjustments based on the preceding year may already be in error in the following year. Conversely, short-term corrections, from only one day or one week of data, may not succeed since adjustments may try to correct daily variations, rather than true trends. We have concluded that a period of one month is the shortest time period to adequately analyze for trend changes,without being too greatly influenced by daily variations.

As an example, suppose we want to correct the predicted ozone levels for the current month, say July, given that we have already seen the observed and the predicted values in June. One correction method, which has been shown to be effective in reducing total errors when applied to the test data (shown in Table 7.1), is given below.

METHOD 1.

1. Record the prediction errors and the predicted values for the past month. Here an error is defined as the difference between the observed and the predicted values. For example, the errors and the predicted values for June can be displayed sequentially, one error followed by one predicted value for each of the thirty days, as shown in Table 7.2.

2. Divide the range of predicted values into subintervals of equal length, such as the intervals, 0 to 4.9 pphm, 5 to 9.9,...,30 to 34.9, and the unbounded interval from 35 and beyond.

3. Using the listing in Table 7.2, group those errors with the predicted values falling in the same subintervals into one class. Figure 7.1 summarizes this procedure.

4. For each row in the above table, rank the numbers from the smallest to the largest in an ascending order, and find the median. Here the median is defined to be:

$$\text{Median} = \begin{cases} \text{the single middle value or} \\ \text{the mean of the two middle values} \end{cases}$$

We shall mark it "M", e.g., if one has nine values, the 5th from either end will be the median. Figure 7.2 shows the ranking and the median for each row previously shown in Figure 7.1.

Very often, a row may contain less than three values, and the median is not well-defined. We recommend that whenever there are less than three values in the 35-∞ row, combine the row with the next highest row, compute

Table 7.1 Test Data Set (Based on Upland, 1977)

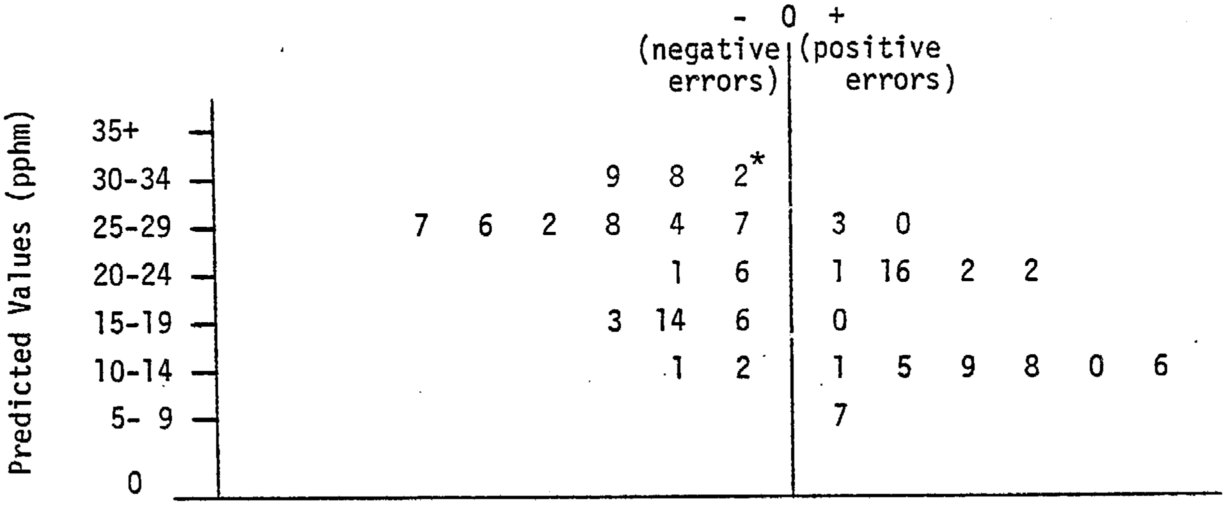
June Predicted (Observed) sequence - 28(31), 24(25), 29(22), 22(38), 28(24), 19(19), 25(17), 19(13), 17(03), 11(09), 12(13), 12(17), 13(12), 09(18), 12(20), 21(23), 21(15), 19(16), 13(13), 08(15), 13(19), 22(24), 26(26), 28(26), 30(28), 27(21), 27(20), 23(22), 31(23), 32(23).

July Predicted (Observed) sequence - 31(20), 21(20), 21(18), 15(06), 19(11), 18(27), 26(24), 31(24), 19(25), 25(14), 27(15), 22(23), 18(23), 21(28), 31(28), 35(29), 33(19), -(18), 22(11), 17(15), 23(31), 33(34), 31(22), 31(24), 28(25), 34(25), 31(23), 34(26), 33(27), 34(18), 29(20).

Table 7.2 Listing of Prediction Errors for June, 1977

Error (Predicted Value) Sequence (PPHM)

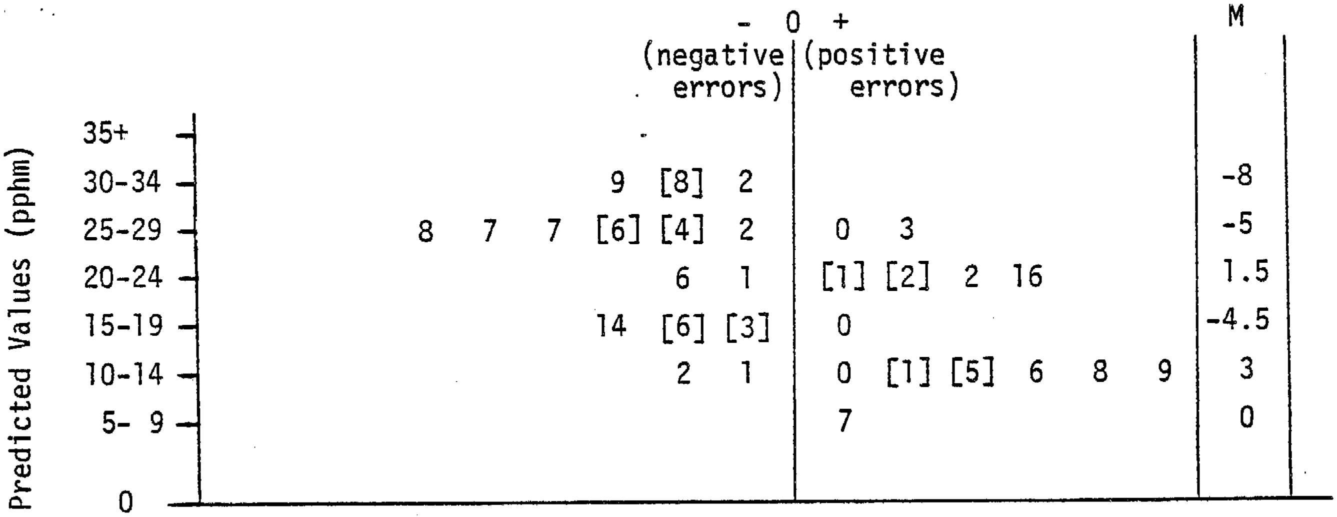
03(28), 01(24), -07(29), 16(22), -04(28), 00(19), -08(25) -06(19), -14(17),-02(11), 01(12), 05(12), -01(13), 09(09) 08(12), 02(21), -06(21), -03(19), 00(13), 07(08), 06(13) 02(22), 00(26), -02(28), -02(30), -06(27), -07(27), -01(23) -08(31), -09(32).



Predicted Values (pphm)												
35+												
30-34				9	8	2*						
25-29	7	6	2	8	4	7	3	0				
20-24					1	6	1	16	2	2		
15-19				3	14	6	0					
10-14					1	2	1	5	9	8	0	6
5- 9							7					
0												

* Numbers shown represent the original errors in pphm.

Figure 7.1 Class Interval Grouping of Prediction Errors



Predicted Values (pphm)													M
35+													
30-34				9	[8]	2							-8
25-29	8	7	7	[6]	[4]	2	0	3					-5
20-24					6	1	[1]	[2]	2	16			1.5
15-19				14	[6]	[3]	0						-4.5
10-14					2	1	0	[1]	[5]	6	8	9	3
5- 9							7						0
0													

Note: This table is good for the July 2 to July 31 period.
(Bracketed numbers indicate values used to compute median, M)

Figure 7.2 Ranked Errors by Prediction Interval

the median, and use that median for the 35-∞ row. For other lower level rows with less than three values, assign the corresponding median the value zero.

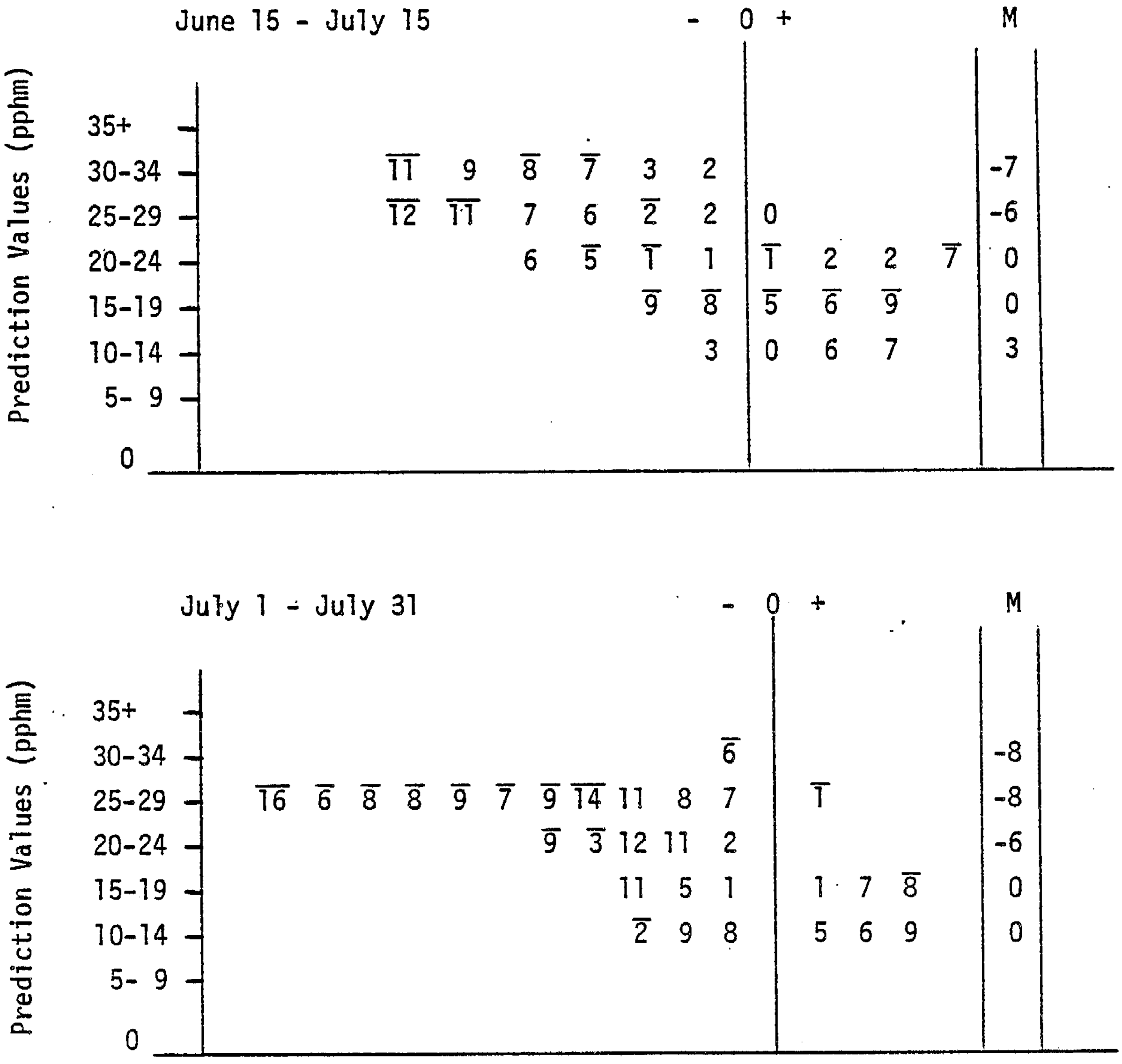
5. In order to make a correction for the predicted value of any one day in the current month, compute the predicted value, say P, from the algorithm and seek for the M value of the row to which P belongs (e.g., from Figure 7.2),and then compute the revised predicted value $\hat{p}$ as P+M. For example, suppose we are interested in the prediction of ozone level for July 14. From the algorithm, we get P = 28 pphm. From Figure 7.2, we see that, for those predicted values falling between 25 and 29, the median error is -5. Thus the revised predicted value is p = 28-5 = 23. Therefore we announce that the predicted value for July 14 is 23 pphm instead of 28 pphm. Figure 7.2 would be valid for the entire month of July. For the August prediction, we need to revise the table using the July data on August 1.

For those rows with $M\leq 2$, no correction is necessary due to the insignificance of the M values.

METHOD 2

The preceding method which requires the user to adjust the error tables once every month, worked fairly well when applied to the test data. Another way of using the method is to adjust the table once every half month. Thus the user may record the prediction errors and the predicted values for the 30 days counted either from the middle of last month or the middle to the current month or from the beginning to the end of the last month depending on the day on which one makes the prediction. (This allows for a more frequent determination of trend changes, but still maintains a one-month record as the basis for making any adjustments.) The rest of the procedure follows exactly the same as in Steps 2, 3, 4 of method 1.

Using the test data as an example, suppose, on July 17, we want to predict the ozone level for July 18. Thus we record the errors and predicted values for the period, June 16 to July 15, and produce the correction table shown in Figure 7.3, according to step 2 in method 1. Since the errors from July 1 to July 15 will be included in the table from



Note: Bars indicate values to be used in subsequent listing.

Figure 7.3 Semi-Monthly Ranking of Prediction Errors

July 1 to July 30, which, in turn,.will be used for the prediction of ozone level from Aug. 2 to Aug. 16, we mark them by a bar,meaning that they will be used again. The table is good for the correction of predictions from July 17 to August 1. For the prediction period of August 2 to August 16,Figure 7.3 would be readjusted by recording the 15 numbers which are marked in the previous table, and adding 15 new numbers. The 15 newly added numbers, again, are marked by bars meaning that they will be used in the construction of the next table. The result is shown in the bottom half of Figure 7.3.

The above two methods together with a method for detecting trends will now be applied to two sets of predicted 1-hour daily maximum ozone levels in UPLA,covering the period from June 1 thorugh October 31, 1977. The results will be compared with the AQMD observed levels for the same period.

Let an error e be defined as the difference between the observed and the predicted values. Let sse = $\sum_i e_i^2$ be the sum of squared errors. Three correction methods will be tested using sse as a measure of performance.

Correction Method 1 - Correct the predicted value according to the median of the errors of the previous calendar month's predictions of that parameter.

Correction Method 2 - Correct the predicted value according to the median of the errors of the previous 30 days' predictions of that parameter. The median is updated every half month.

Correction Method 3 - Correct the predicted value according to the median of the errors for a given observed level in the previous 30 days' forecasting, again adjusting the median once every half month.

Let

$s\hat{s}e_{(i)}$ = sum of squared errors after correction using method i, where i = 1,2, and 3.

$$r_i\% = \frac{(sse-s\hat{s}e_{(i)})}{sse} \times 100\% = \text{percent of total errors reduced by using method i.}$$

Results are shown in Table 7.3.

Table 7.3 Test Result

Test Set No. 1	sse	$s\hat{s}e_{(1)}$	$r_{(1)}\%$	$s\hat{s}e_{(2)}$	$r_{(2)}\%$	$s\hat{s}e_{(3)}$	$r_{(3)}\%$
June	0.0705						
July	0.1663	0.1317	20.8%	0.1050	36.8%	0.0981	41.0%
August	0.2399	0.1561	34.9%	0.1784	25.6%	0.2053	14.4%
Sept.*	0.1298	0.0936	27.8%	0.0896	30.9%	0.1274	1.8%
Oct.	0.1279	0.1081	15.4%	0.1115	12.8%	0.1163	9.0%
Test Set No. 2							
June	0.1180						
July	0.1914	0.0997	47.9%	0.0999	47.8%	0.1952	-1.9%
August	0.2668	0.1972	26.0%	0.1896	28.9%	0.2626	1.5%
Sept.	0.1572	0.1273	**19.2%	0.1654	**16.22%		
Oct.	0.1461	0.1469	**-0.54%	No Correction	**		

*Only 3 weeks' data available.

**From the tables covering the period from June 1 through August 31, we observed that the M values for rows below 0.25 ppm were zero most of the time. Thus for September, we decided that no correction was needed for any prediction value below 0.25 ppm.

Intuitively, Method 3 seems reasonable since the error distribution of a prediction algorithm which systematically underpredicts (or overpredicts) the pollutant level should present a positive median (or negative median). Accordingly the median should be subtracted (or added) from the predicted value assumption from the algorithm. Unfortunately, when tested on real data, this "reasonable" method does not work. Nevertheless, the test result of this method provides one with an impression on the relative merit of Method 1 and Method 2.

From the test result, one sees that, overall, Method 1 and 2 achieve at least 20% of reduction in total errors. During October the original algorithm performs better (smaller sse value) than during the rest of months in the set, resulting in lower reductions in total error by Method 1 and Method 2.

After the application of Methods 1 and 2 to the test sets, it can be seen that, for small predicted values (say below 0.25 ppm), no correction is necessary. Thus one only needs to correct large predicted values according to the table.